Meta-Analysis Techniques in Medical Research: A Statistical Perspective

Rebecca Jane Hardy

A Thesis Submitted for the Degree

of Doctor of Philosophy

London School of Hygiene and Tropical Medicine

April 16, 1995

Abstract

Meta-analysis is now commonly used in medical research. However there are statistical issues relating to the subject that require investigation and some are considered here, from both a methodological and a practical perspective.

Each of the fixed effect and the random effects models for meta-analysis are based on certain assumptions and the validity of these is investigated. A formal test of the homogeneity assumption made in the fixed effect model may be performed. Since the test has low power, simulation was used to investigate the power under various conditions. The random effects model incorporates a between-study component of variance into the model. A likelihood based method was used to obtain a confidence interval for this variance and also to provide an interval for the overall treatment effect which takes into account the fact that the between-study variance is estimated, rather than assuming it to be known.

In order to obtain confidence intervals for the treatment effect for both the fixed effect and the random effects models, distributional assumptions of normality are usually made. Such assumptions may be checked using q-q plots of the residuals obtained for each trial in the meta-analysis. In both meta-analysis models it is assumed that the weight allocated to each study is known, when in fact it must be estimated from the data. The effect of estimating the weights on the overall treatment effect estimate, its confidence intervals, the between-study variance estimate and the test statistic for homogeneity, is investigated by both analytic and simulation methods.

It is shown how meta-analysis methods may be used to analyse multicentre trials of a paired cluster randomised design. Meta-analysis techniques are found to be preferable to previously published methods specifically developed for the analysis of such designs, which produce biased and potentially misleading results when a large treatment effect is present.

Contents

List of Tables

List of Figures

Acknowledgements

I wish to thank my supervisor Simon Thompson for his guidance and encouragement throughout the undertaking of the research contained in this thesis.

I am grateful to the Medical Research Council for providing my funding. Furthermore, I am indebted to Professor D Wood for allowing me to use data from the British family heart study, and also to Professor TW Meade and PJ Brennan for allowing me to use data from the MRC trial of the treatment of mild hypertension.

I also wish to thank Professor H van Houwelingen for providing a computer program for part of the research and Professor A Donner for helpful comments regarding the analysis of paired cluster randomised trials.

1 Existing Statistical Methods in Meta-Analysis

Chapter 1 provides an introduction to standard meta-analysis techniques and sets the scene for the subsequent research described in later chapters. Section 1.1 contains a review of the current meta-analysis literature and provides general background information. Section 1.2 outlines the structure and the aims of the thesis, while Section 1.3 introduces two data sets which are used in the thesis as examples. The standard meta-analysis methods are then introduced, with Section 1.4 considering the issue of hypothesis testing and the next three sections that of estimation. Section 1.5 describes the different fixed effect methods of meta-analysis, Section 1.6 considers the issue of heterogeneity and Section 1.7 outlines the random effects method of meta-analysis. The question of how to display meta-analysis data is addressed in Section 1.8 and finally, Section 1.9 contains a discussion and comparison of the two different meta-analysis approaches, those of the fixed effect and the random effects models.

1.1 Background

Meta-analysis can be defined as the statistical evaluation of a collection of analytic results for the purpose of integrating the findings [1]. Researchers in psychology and education were the first, in the 1970s, to define meta-analysis and begin to develop the statistical methodology. Meta-analyses were, however, rare in the medical literature until the early 1980s, but have proliferated in the last few years [2], although one of the first medical meta-analyses was performed as far back as 1977 [3].

The aim of such an analysis is "to obtain information that cannot be ascertained from any of the studies alone" [4]. Peto discussed the importance of meta-analysis, indicating that while moderate differences in mortality rates may be humanly

worthwhile, in many circumstances it is very difficult to detect a 10% - or even 20% - reduction in risk of death [5]. Studies involving at least 1000 deaths may often be required to detect such effects reliably. Lack of money, resources and time may limit the size of a single trial, and so while it is generally emphasized that meta-analysis should never be a substitute for the single, large well-designed study [6, 7, 8], it is clear that the combining of results from different trials is a desirable and necessary technique in the field of medical research. However, there is considerable debate as to when and how the data should be combined in a formal manner.

In the past, clinicians have relied heavily on narrative reviews of literature to define the current state of knowledge on any particular therapy. However, it is common for similar trials on the same treatment to produce apparently conflicting results. The situation becomes even more confused when these trials differ in terms of the treatment regimen, treatment duration and patient characteristics. Hence, interpretation of all the information available is difficult and the conclusions reached will be highly subjective and may depend greatly on the perspective of the individual reviewer. Indeed, such reviews have been criticised as being haphazard and biased [9]. Meta-analysis can simply be thought of as a more structured approach to this traditional literature review which attempts to produce an objective measure of the overall benefit of the therapy being considered [10]. Nevertheless, there is still scope for differing conclusions to be reached.

Chalmers et al. [11] carried out a study looking at the reproducibility of meta-analysis and found that there were cases where meta-analyses on the same therapy arrived at different conclusions. However, encouragement may be gained from the fact that these observed disagreements were usually in terms of degree rather than direction. A recent example in the literature has been the disagreement regarding meta-analyses of the trials of serum cholesterol reduction [12]. Such discrepancies may arise owing to different investigators including a different collection of studies in

their meta-analysis because of, for example, different literature search strategies or different inclusion criteria. A great deal of attention has been focused on the problem of which studies to include in a meta-analysis and Naylor [13] holds the opinion that "methodology is less important than determining which results are to be aggregated". It is generally agreed that there is a need for scientific rigour throughout the whole meta-analysis process, including the initial literature search [6, 14]. The use of a meta-analysis protocol specifying all procedures, especially those related to the selection of trials, has been advocated [6]. The possibility of "publication bias" [15], whereby published studies differ systematically from unpublished ones is a widely recognised problem in meta-analysis [14]. Hence, it is advised that efforts should be made to minimise this potential bias by tracking down relevant unpublished material [10]. Furthermore, guidelines have been proposed which were designed to help to minimise bias in meta-analyses [2]. Meta-analyses usually combine only the information directly available from the literature and so are reliant on the validity of the analysis in the original trials. Hence, it has recently been advocated that meta-analyses should be based on the reanalysis of individual patient data, as this provides the least biased and most reliable results [16].

Considerable research has also been carried out on how the quality of each study included in a meta-analysis may influence the results [17, 18, 19] and, furthermore, of ways to incorporate this quality assessment into the statistical analysis [17]. This may involve the use of a specific quality score threshold meaning that poor quality studies are excluded, or the incorporation of the quality scores as weights [20]. Greenland [21] however, describes quality scoring as "the most insidious form of subjectivity masquerading as objectivity" and states that it can obscure important sources of heterogeneity.

Thacker [22] was concerned that using sophisticated meta-analysis techniques could lead to "unwarranted comfort with one's conclusions" if the initial data used is

of a poor standard. However, O'Rourke [14] views meta-analysis as an ideal means for uncovering and correcting inadequacies in previous research. It has even been suggested [8] that all clinical trials should be started with the notion of meta-analysis in mind in order to help future reviewers. In this way, evidence in the form of a meta-analysis would keep accumulating as the results of each new study became available [23]. Specifically, it has been shown that continuously updated literature reviews, as exemplified by the Oxford Database of Perinatal Trials, can shorten the time between research discoveries and clinical implementation of treatment strategies [24]. This database has been extended to include results of meta-analyses and will be updated as each relevant new trial is published. As well as the formation of databases, there is also a move to generate user-friendly meta-analysis computer packages [25], such as that produced by the Cochrane Collaboration [26]. These developments will allow more meta-analyses to be carried out, but it should be emphasised that care and thought should still go into each analysis and data should never be fed blindly into a program to obtain a result [25]. There is also the need for emphasis in computer software on the issue of heterogeneity and how it should be investigated [27].

Issues relating to the identification, selection and quality of trials for a meta-analysis have been widely discussed in the literature and will not be pursued in this thesis. Concentration is, instead, focused on the methodological issues relating to the statistical methods used in meta-analysis. The majority of the work on meta-analysis in the medical field has been with regards to randomised clinical trials, and this is where the emphasis will lie here. There has been some work relating to epidemiological studies [28, 29], but the results of such studies are even more problematic to combine than those of clinical trials owing to the additional variations in design and the greater scope for the existence of biases in individual studies.

Many of the statistical techniques applied to meta-analysis have been long established for the purpose of combining various forms of experimental data. It is

only much more recently that these methods have been applied to the problem of combining the results of completely separate studies. For example, methodology has been extracted from the work of Cochran in the 1950s [30] relating to the combination of estimates from different experiments. The statistical aspect of meta-analysis can be split into two components; hypothesis testing and estimation. In the case of hypothesis testing, the null hypothesis is that all individual studies in the meta-analysis have in truth zero treatment effect. Estimation deals with the calculation of an overall treatment effect together with a relevant confidence interval. The issue of estimation is the more problematic of the two and is therefore considered in greater detail in this thesis. Most meta-analyses in the medical field have concentrated on estimating and testing the common treatment effect, which is assumed to be equal for all trials [31]. This is the fixed effect approach, which makes the assumption that the true treatment effect is the same in all of the individual studies included in the meta-analysis, that is the treatment effects are homogeneous. Several estimates have been proposed and these are described in Section 1.5.

It is nearly always unreasonable to assume homogeneity in medical contexts, and the combining of heterogeneous material is a commonly cited threat to the validity of meta-analysis [32]. There has been concern about the practice of merely publishing the numerical results of a fixed effect analysis, where "little attention is paid to possible heterogeneity in effect sizes between trials" [31]. Thompson and Pocock [33] stress that a single weighted average of the separate treatment effects is difficult to interpret, as it is not clear as to what treatment or what population of patients it applies. This interpretational problem means that clinicians may find it difficult to apply the results of a meta-analysis to a practical situation, since they must decide whether the results are generalisable to their own specific case [34]. On the other hand, it has been argued [13] that since a broader range of patients and practices has been incorporated into the data, the generalisability of the results from the combination of several small trials may be superior to that of a single large trial. Hence, it is possible

to argue that the results of a meta-analysis of small trials are more applicable to the practice of medicine, in which the patients encountered are rarely homogeneous [32].

Various formal tests of heterogeneity exist (Section 1.6), but they lack power and so a negative result should not be interpreted as implying that the treatment effects are homogeneous [35]. Therefore, L'Abbé et. al. [10] suggest that in the case of a non-significant result, the investigators should "resort to informed judgement and examine a graphic display of heterogeneity...". It has generally been stressed in the literature that the investigation of heterogeneity and its possible causes are of the utmost importance when carrying out a meta-analysis. An investigation of subgroups of the studies can be carried out whereby further important questions, such as for whom and under what circumstances the treatment works best, may be considered [10]. In fact the ability to explore such questions can be viewed as a great advantage of meta-analysis [36]. However, care should be taken when such an investigation of heterogeneity is undertaken, since such investigations will tend to be post-hoc and hence the problems are similar to those which occur when undertaking subgroup analyses in a single clinical trial [27, 37]. Furthermore, it is generally not a simple matter to isolate a single source of heterogeneity, and it may be that a number of possibilities exist or there may be no explanation apparent at all.

DerSimonian and Laird [38] proposed an alternative to the fixed effect approach which does take account of the between-study variation in the true treatment effects. This is known as the random effects model (Section 1.7), since it incorporates a between-study component of variance. The random effects model allows for the extra uncertainty in a set of heterogeneous data and produces an overall estimate of treatment effect with a suitably widened confidence interval. However, the model has been criticised as unrealistic for making the assumption that the trials included in the meta-analysis are a random sample taken from some hypothetical universe of trials. It has also been criticised as being an easy option and an excuse for not investigating

the causes of heterogeneity fully [21].

As well as these standard approaches to meta-analysis, further methods have recently been developed. The use of empirical Bayes methods for meta-analysis has been proposed [31, 39], whereby shrunken estimates of the treatment effect for each trial are obtained. These empirical Bayes estimates incorporate the information from the full set of data in order to provide a more precise estimate for each trial. A fully Bayesian approach has been adopted by other researchers. Carlin [40] and Skene and Wakefield [41] based methods on three-stage hierarchical models and then used Monte-Carlo type methods in order to obtain the solutions. Malec and Sedransk [42] used a prior distribution which reflected the belief that there are subsets of trials in the data such that within each subset each trial produces a similar result. The composition of these subsets was, however, considered to be uncertain. In addition, Eddy, Hasselblad and Shachter [43] proposed a Bayesian approach to meta-analysis which they called the "confidence profile method".

An alternative development in meta-analysis has been the use of likelihood theory. Goodman [44] produced plots of the "support curves" in order to obtain the parameter values which received the most "support" from the available data. Within this likelihood framework he proposed both a fixed effect and a random effects model. A further likelihood based approach, for binary outcomes, proposed by van Houwelingen et al. [45] uses the exact conditional distribution of each 2x2 table. The likelihood approach to meta-analysis will be considered further in Chapter 2.

1.2 Aims and Objectives of the Thesis

Meta-analysis, as an objective way of reviewing research, is now firmly established in the area of medical research. However, there is still much debate as to the best statistical approach to the analysis and, furthermore, there are unresolved statistical

problems relating to both the standard fixed effect and random effects models which require investigation. This thesis addresses a number of these statistical issues from both a methodological and a practical perspective.

The standard fixed effect and random effects models are each considered critically and are compared with each other and with more novel methods, using illustrative practical examples. It is accepted that neither model forms a realistic basis for an overall estimate of treatment effect, as the assumptions underlying each cannot be met. The validity of certain of these assumptions are investigated and the robustness to deviations from them are discussed.

A further aim of the research is to extend the use of meta-analysis methods to other types of data. It is shown how meta-analysis models may be useful in the analysis of multicentre trials where there is heterogeneity between centres. It is also shown how the random effects meta-analysis model can be used to analyse a single trial which has a paired cluster randomised design.

Although the basis of much of the research is methodological, the practical implications of the findings are always discussed and practical examples used where possible to illustrate the points. The main data set used to illustrate the analysis of a paired cluster randomised design is also used to exemplify the techniques in practice and in order to pursue a practical data set in greater detail. Two other data sets are used regularly as examples and these are introduced in Section 1.3 and will be referred to at various stages throughout the thesis.

The rest of the current Chapter reviews the present state of statistical methods for meta-analysis in medical research. The standard methods for both testing and estimation are described and numerical examples are used to compare the results obtained. Chapter 2 extends the estimation ideas presented in the introduction, focusing on the random effects model and particularly on the issue of the estimation

of the between-study component of variance. An approximate likelihood method is proposed which produces both a confidence interval for the between-study variance and also a confidence interval for the overall treatment effect which takes into account the fact that the between-study variance is estimated. This approximate method, based on the marginal likelihood, is found to be very comparable to a method based on the full likelihood for binary data under most circumstances.

Both the standard fixed effect model and random effects model must make distributional assumptions of normality if confidence intervals are to be obtained. Chapter 3 proposes the use of q-q plots of the residuals, obtained for each trial in the meta-analysis, in order to investigate these assumptions. The issue of testing for normality is also considered.

Chapter 4 considers the power of the test for heterogeneity, which is known to be low. The test is investigated using simulation in an attempt to quantify the power and to identify situations where the power will be particularly poor. The standard test is also compared with an alternative, supposedly more powerful test. Chapter 5 extends the simulation work of Chapter 4 to investigate the effect that estimating the individual within-study variances has on the power of the test, since the null distribution of the test statistic is conditional on the assumption that they are known. The work is then developed to look at the effect of this estimation on the overall fixed effect and random effects estimates and their confidence intervals. Analytic work is carried out to try to obtain improved estimates which allow, at least to some extent, for the estimation of the weights.

Chapters 6 and 7 deal with the analysis of paired cluster randomised trials. Chapter 6 describes a more detailed analysis of a single data set, namely the British family heart study, thus illustrating how the ideas and methods previously described are useful in practice. It also provides an opportunity for an investigation into and a discussion of sources of heterogeneity as well as a chance to consider the analy-

sis of multiple endpoints in meta-analysis. Chapter 7 compares the meta-analysis techniques, for both testing and estimation, with previously published methods for paired cluster randomised trials. Meta-analysis estimation techniques were found to be preferable to those developed specifically for the analysis of paired cluster randomised designs, which were shown to produce biased results. Finally, the findings are summarised and the overall conclusions drawn in Chapter 8.

1.3 Introduction to Data Sets

1.3.1 A meta-analysis of nine clinical trials looking at the effect of taking diuretics during pregnancy

A meta-analysis of nine randomised controlled clinical trials which was published by Collins, Yusuf and Peto [46] is used throughout the thesis as an example. One aim of this meta-analysis was to look at the effect of diuretics during pregnancy on the incidence of pre-eclampsia. The term pre-eclampsia is used to describe the development of hypertension with proteinuria or oedema, or both, during pregnancy. Pre-eclampsia is known to increase the risk of a perinatal death, which is the outcome of ultimate importance. Perinatal mortality is, however, a difficult outcome to study in a clinical trial as only a few pregnancies end in a death and only a few of these are associated with pre-eclampsia [46]. Since perinatal death is such a rare occurrence, all the single clinical trials looking at this issue have been too small to detect any differences in mortality and so have tended to concentrate on the effect of the treatment on pre-eclampsia. Even using a meta-analysis, there were still too few deaths to achieve adequate power to detect any treatment effect. Hence, the outcome generally used in this thesis, when looking at the diuretics trials data, is the presence or absence of pre-eclampsia, although the number of stillbirths is considered in one instance.

Reliable data were available for nine out of the eleven trials published on diuretics during pregnancy since 1960, and the meta-analysis of these nine trials lead to a total sample size of 7 000 women and over 600 cases of pre-eclampsia (Table 1). It can be seen that the number of women and the number of cases vary considerably across the trials. Furthermore, the individual study odds ratios vary with six of the trials showing a positive effect of diuretics, but the other three showing an adverse effect. The variation in estimates of treatment effect is not surprising, since the trials differed from each other in many respects. The entry criteria for patients varied, as did the treatment regimens and the definition of pre-eclampsia. Some trials also had greater problems with withdrawals and non-compliance.

Table 1: Results and odds ratios for the nine trials included in the meta-analysis looking at the effects of diuretics on the occurrence of pre-eclampsia during pregnancy

Trial number	First author of paper	Cases of pre-eclampsia/Total number of patients		Odds Ratio
		Treated	Control	
1	Weseley	14/131(10.7%)	14/136(10.3%)	1.04
2	Flowers	21/385(5.5%)	17/134(12.7%)	0.40
3	Menzies	14/57(24.6%)	24/48(50.0%)	0.33
4	Fallis	6/38(15.8%)	18/40(45.0%)	0.23
5	Cuadros	12/1011(1.2%)	35/760(4.6%)	0.25
6	Landesman	138/1370(10.1%)	175/1336(13.1%)	0.74
7	Krans	15/506(3.0%)	20/524(3.8%)	0.77
8	Tervila	6/108(5.6%)	2/103(1.9%)	2.97
9	Campbell	65/153(42.5%)	40/102(39.2%)	1.14

1.3.2 A multicentre trial looking at the treatment of mild hypertension

Meta-analysis methods can also be used in the analysis of multicentre clinical trials, as each centre can be considered as being equivalent to a separate trial. Although the protocol followed should be the same in each centre, meaning therefore that there is less scope for clinical heterogeneity to exist, there may still be differences due to geographical location and demographic characteristics of the patients. There may also be differences in the centres' interpretation of the protocol, the additional care given and the skill with which the treatment is administered. Hence, the possibility of between-centre differences should at least be considered in the analysis of a multicentre trial.

The main aim of the Medical Research Council (MRC) mild hypertension trial was to determine whether drug treatment of mild hypertension (phase V diastolic pressure 90-109 mmHg) reduces the rates of stroke, of death due to hypertension and of coronary events in men and women aged 35-64 years [47]. A subsidiary aim was to compare the blood pressure in two groups of patients on active treatment, those taking bendrofluazide and those taking propranolol. The outcome considered most often in this thesis is the blood pressure reduction (both diastolic and systolic) over the first year of the trial in the combined treatment group compared to the control group, as this provides the opportunity to analyse a continuous outcome measure. The measurement of treatment effect used is a difference in two means, that is the difference between the mean reduction in blood pressure in the treatment group and the mean reduction in blood pressure in the control group.

In total, 17 354 patients were randomly allocated at entry to the trial to take bendrofluazide or propranolol or placebo tablets. The patients were recruited from 190 centres (mostly general practices) distributed throughout England, Scotland and Wales. The original analysis [47] made no allowance for possible differences across these centres, analysing the whole set of data without stratification for centre. All

analyses are by 'intention to treat', that is patients were analysed as belonging to the group to which they were randomised, irrespective of what treatment they actually received.

The number of patients in each centre varied greatly, with some centres recruiting less than 10 patients and others recruiting over 100. Most analyses had to be carried out using 189, rather than 190, centres, since centre number 1 recruited only one patient to the placebo group, meaning that a variance for the reduction in blood pressure could not be calculated for that centre. The total number of strokes observed was 169 and the total number of coronary events was 456. The overall average reduction in diastolic blood pressure over the first year in the treatment group was 11.7mmHg, while for systolic blood pressure it was 23.7mmHg. The average blood pressure was also reduced over the trial period in the placebo group. Diastolic blood pressure was reduced on average by 6.6mmHg and systolic blood pressure by 13.3mmHg. This may be as a result of the so called 'placebo effect'.

1.4 Hypothesis Tests in Meta-Analysis

In a meta-analysis of k trials, a hypothesis test may be carried out in order to see whether the k differences from a zero treatment effect observed are greater than would be expected by chance. The null hypothesis is, therefore, that the true treatment effect θ_i in each trial i, $i = 1, ..., k$, is equal to zero, that is $H_0 : \theta_1 = ... = \theta_k = 0$. Hence, if there is sufficient evidence that any one of the individual trial estimates deviates from zero, then the null hypothesis will be rejected.

If it is assumed, at least asymptotically, that the estimate of θ_i has a normal distribution with mean θ_i and variance v_i then under the null hypothesis, this estimate $\hat{\theta}_i$ has a normal distribution with zero mean and variance v_i. When the interest lies in the absolute values of the departure of the θ_i from no treatment effect, an

appropriate test is based on the squares of the individual observed treatment effects, thus not accounting for the direction of the effect. Standardising these squared effects by dividing by the variance v_i, means that $\hat{\theta}_i^2/v_i = w_i\hat{\theta}_i^2$, where $w_i = 1/v_i$, has a χ_1^2 distribution if θ_i=0 [12]. Hence, assuming that all trials are independent, summing over all k trials gives a test statistic for H_0

$$\sum_{i=1}^{k} w_i\hat{\theta}_i^{\,2} \tag{1}$$

which has a χ_k^2 distribution under H_0. However, the use of the squares of the treatment effects means that this test is 'general' in that it has no specific alternative hypothesis against which it is particularly powerful [12]. The general alternative hypothesis is $H_1 : \theta_i \neq 0$ for at least one i, $i = 1, ..., k$. This test therefore lacks power against certain alternative hypotheses of particular interest, such as $H_1 : \theta_i < 0$ (or $\theta_i > 0$) for all i.

One appropriate test which is powerful against these directional alternatives is again based on the asymptotic normality of each trial estimate, but does account for the direction of each estimate. For each study, if θ_i=0, $w_i\hat{\theta}_i$ has a normal distribution with zero mean and variance given by $1/v_i = w_i$ [4]. Hence, for the null hypothesis, $H_0 : \theta_1 = ... = \theta_k = 0$, the sum $\sum_{i=1}^{k} w_i\hat{\theta}_i$ has a normal distribution with zero mean and variance $\sum_{i=1}^{k} w_i$ [4]. Therefore the test statistic

$$\frac{\sum_{i=1}^{k} w_i\hat{\theta}_i}{\sqrt{\sum_{i=1}^{k} w_i}} \tag{2}$$

has a standard normal distribution under the null hypothesis [48]. Equivalently, the square of the statistic given in (2) follows a χ_1^2 distribution [4, 12]. The rejection of the null hypothesis can still, however, only be interpreted as being evidence that at least one treatment effect is different from zero.

The Mantel-Haenszel test is a particular example of a test, although not directly equivalent to (2), which is powerful against the directional alternative hypotheses [49]. It can only be used in situations where the outcome measure is binary, unlike test statistics (1) and (2) which may be used for both binary and continuous data. The null hypothesis is, once more, that the treatment effect in every study is zero. Then, under H_0, and using the notation given in Table 2, for each trial i, $i = 1, ..., k$, the number of observed events in the treatment group a_i, conditional on the total number of patients in the treatment group $a_i + b_i = n_{i1}$, has a hypergeometric distribution with a mean given by $n_{i1}m_{i1}/N_i$ and a variance of $n_{i1}n_{i2}m_{i1}m_{i2}/N_i^2(N_i - 1)$ where $N_i = n_{i1} + n_{i2}$, $m_{i1} = a_i + c_i$ and $m_{i2} = b_i + d_i$ (Table 2). Furthermore, since in a meta-analysis, the strata are k independent trials, the total observed number of events $\sum_{i=1}^{k} a_i$, simply has a mean of $\sum_{i=1}^{k} E(a_i)$ and a variance of $\sum_{i=1}^{k} var(a_i)$. The variance of the sum of the differences between the observed and the expected number of events $\sum_{i=1}^{k}(a_i - E(a_i))$ is therefore equal to the sum of the individual variances of the $(a_i - E(a_i))$ terms [50]. Hence, the test statistic is given by

$$\frac{[\sum_{i=1}^{k}(a_i - E(a_i))]^2}{\sum_{i=1}^{k} var(a_i)} \tag{3}$$

which, it may be shown, has an asymptotic χ_1^2 distribution.

A continuity correction for (3) may be necessary, particularly when the numbers involved are small and hence the test becomes

$$\frac{[|\sum_{i=1}^{k}(a_i - E(a_i))| - 0.5]^2}{\sum_{i=1}^{k} var(a_i)} \tag{4}$$

A test which is identical to the Mantel-Haenszel test, but which uses different notation, is known as Peto's test [51]. Peto considers the differences between the observed O_i and the expected E_i number of events in the treatment group. The

Table 2: Notation for the frequency table of the results of the i^{th} study from which the odds ratio is calculated

Number of patients		Event		Total
		Yes	No	
Group	Treatment	a_i	b_i	$a_i + b_i = n_{i1}$
	Control	c_i	d_i	$c_i + d_i = n_{i2}$
Total		$a_i + c_i = m_{i1}$	$b_i + d_i = m_{i2}$	N_i

sum of these differences is squared and divided by the total variance under the null hypothesis V_i and hence the statistic is written,

$$\frac{[\sum_{i=1}^{k}(O_i - E_i)]^2}{\sum_{i=1}^{k} V_i} \tag{5}$$

However, by noting that $O_i = a_i$, $E_i = n_{i1}m_{i1}/N_i$ and $V_i = n_{i1}n_{i2}m_{i1}m_{i2}/N_i^2(N_i - 1)$ it can be seen that this test is exactly the same as the Mantel-Haenszel test.

Examples which illustrate the use of the above tests are presented in Section 1.5.5, after the standard methods of estimation in meta-analysis have been described.

1.5 Fixed Effect Methods

The principle for estimating an overall treatment effect is that the observed treatment effect within each trial should be averaged over all trials [36]. In order to carry out such a procedure, however, assumptions are required. Initially the main issue in the fixed effect approach, where homogeneity of treatment effects across studies, $\theta_1 = \theta_2 = ... = \theta_k = \theta$, is assumed, is how to weight the individual trial estimates.

The general idea is to give the greatest weight to those studies with the most precise individual estimates $\hat{\theta}_i$, $i = 1, .., k$, and then the weighted average, given by Woolf [52], takes the general form,

$$\hat{\theta}_f = \frac{\sum_{i=1}^{k} w_i \hat{\theta}_i}{\sum_{i=1}^{k} w_i} \tag{6}$$

where w_i is the weight associated with trial i. If each $\hat{\theta}_i$ is an estimate of a common θ, then the expected value of the weighted average, assuming that the weights are fixed constants (although the effects of relaxing this assumption will be considered in Chapter 5) will be,

$$\frac{\sum_{i=1}^{k} w_i}{\sum_{i=1}^{k} w_i} E(\hat{\theta}_i) = \theta \frac{\sum_{i=1}^{k} w_i}{\sum_{i=1}^{k} w_i} = \theta \tag{7}$$

This means that any choice of w_i, $i = 1, ..., k$, will lead to an unbiased estimate of the true treatment effect. However, the most precise estimate of the overall treatment effect θ, that is the one with the minimum variance, is obtained by calculating the weighted average (6) and taking the weight for the i^{th} study to be the inverse of the variance v_i, that is w_i=$1/v_i$ [53].

The general estimate given in (6) may be applied to various outcome measures. For example, the difference in means between a treatment and a control group could be the measure used if a continuous outcome measure were to be analysed. However, the odds ratio is used here as a convenient measure to consider, as a comparison may then be made of several estimation methods. Four methods of estimating an overall treatment effect are now described in Sections 1.5.1 to 1.5.4. These methods, together with the hypothesis tests described in Section 1.4, are then illustrated with an example in Section 1.5.5.

1.5.1 Inverse-variance method

It follows, that by weighting according to the inverse of the variance, larger trials which have estimates with smaller variances are given most weight, while small studies with large variances are given less weight. It is actually the logarithm of the odds ratio that is taken as θ here, since such a transformation improves the normality of the distribution of estimated treatment effects. Hence, taking the log odds ratio $ln(a_i d_i / b_i c_i)$ (Table 2) as $\hat{\theta}_i$, the variance v_i can then be estimated by $(1/a_i)+(1/b_i)+(1/c_i)+(1/d_i)$ [50].

Furthermore, assuming that $\hat{\theta}_i$ is approximately distributed as $N(\theta, v_i)$ and that the w_i=$1/v_i$ are known, then the variance of the overall log odds ratio is $1/\sum_{i=1}^{k} w_i$. This allows a 95% confidence interval for the estimate to be obtained,

$$\hat{\theta}_f \pm 1.96\sqrt{var(\hat{\theta}_f)} \tag{8}$$

The estimate of the overall odds ratio and its corresponding confidence interval may then be found by exponentiating the relevant values calculated for the log odds ratio.

1.5.2 Mantel-Haenszel method

The Mantel-Haenszel estimate [49] is well established and much used for the purpose of combining information across a set of 2x2 tables. The result can be used in the context of a meta-analysis by considering each study as a separate stratum producing an individual estimate of the odds ratio $a_i d_i / b_i c_i$. A weighted average is again used, but the weight w_i in stratum i is now taken to be $b_i c_i / N_i$ (Table 2). These weights are approximately inversely proportional to the variance of the log odds ratio under the condition that the stratum-specific odds ratios are near unity [53], thus implying that $a_i d_i = b_i c_i$. By substituting the relevant values, $\hat{\theta}_i$=$a_i d_i / b_i c_i$ and $w_i = b_i c_i / N_i$,

into the general formula (6), the overall Mantel-Haenszel estimate of the odds ratio is obtained:

$$\hat{OR}_{MH} = \frac{\sum_{i=1}^{k} a_i d_i / N_i}{\sum_{i=1}^{k} b_i c_i / N_i} \tag{9}$$

Robins, Breslow and Greenland [54] obtained an approximation to the variance of the logarithm of the Mantel-Haenszel estimate. By letting $P_i = (a_i + d_i)/N_i$, $Q_i = (b_i + c_i)/N_i$, $R_i = a_i d_i / N_i$ and $S_i = b_i c_i / N_i$, the variance is given by the formula,

$$var(log\hat{OR}_{MH}) = \frac{\sum_{i=1}^{k} P_i R_i}{2(\sum_{i=1}^{k} R_i)^2} + \frac{\sum_{i=1}^{k} (P_i S_i + Q_i R_i)}{2 \sum_{i=1}^{k} R_i \sum_{i=1}^{k} S_i} + \frac{\sum_{i=1}^{k} Q_i S_i}{2(\sum_{i=1}^{k} S_i)^2} \tag{10}$$

This variance may then be used to calculate the confidence interval for the overall odds ratio in a similar way to that shown in (8).

1.5.3 Peto method

Peto's method [5, 51], which is the estimation procedure related to the Peto test described in Section 1.4, produces an approximate estimate of the overall odds ratio through a comparison of the number of events observed with the number expected under the null hypothesis of no treatment effect in each particular trial. For each trial the observed minus the expected number of events $(O_i - E_i)$ is calculated, where $O_i = a_i$ and $E_i = n_{i1} m_{i1} / N_i$ (Table 2). Peto's argument [5] for using this measure is that if there were no treatment effect, then $(O_i - E_i)$ for each trial would have an equal chance of being either negative or positive, and hence the total of all the $(O_i - E_i)$ would be close to zero. However, if a beneficial treatment effect were present then the $(O_i - E_i)$ terms would tend to be negative (since O_i would tend to be less than E_i) and, although this trend may not be noticeable in individual studies, it may stand

out when the total is calculated. Furthermore, by dividing $(O_i - E_i)$ by its variance, a good approximation to the log odds ratio results, providing the odds ratio is not too far away from unity [55]. This stems from the fact that such an approximation is the first Newton-Raphson step from zero towards the maximum likelihood estimate [51, 56] and in general a parameter may be estimated by Z_i/V_i, where Z_i is the efficient score and V_i is Fisher's information [4]. In this specific situation, $(O_i - E_i)$ is the efficient score statistic for the log odds ratio and its variance V_i, given by $V_i = n_{i1}n_{i2}m_{i1}m_{i2}/N_i^2(N_i - 1)$ (Table 2), is Fisher's information [4]. The overall log odds ratio can then be estimated by simply adding up the differences $(O_i - E_i)$ and dividing the resulting total by the sum of the individual variances V_i, to give

$$\hat{logOR_P} = \frac{\sum_{i=1}^{k}(O_i - E_i)}{\sum_{i=1}^{k} V_i} \tag{11}$$

It can also be shown that by taking $\hat{\theta}_i$ equal to $(O_i - E_i)/V_i$ and w_i equal to be $V_i = 1/var(\hat{\theta}_i)$ and substituting into the formula for the general weighted mean (6), the same overall estimate as shown in equation (11) is obtained. The variance of this estimator is, therefore, again given by $1/\sum_{i=1}^{k} w_i$, which in this case is equal to $1/\sum_{i=1}^{k} V_i$. A confidence interval may be calculated in the usual way using this variance term.

1.5.4 Logistic regression

Logistic regression may be used to carry out a meta-analysis where the outcome measure is an odds ratio [28, 57]. A variable representing 'trial' is treated as a factor with k levels, so that each level of the factor corresponds to an individual trial in the meta-analysis. After including such a factor in the model, together with a second factored variable with two levels representing treatment group, the required log odds ratio of treated patients compared to control patients is obtained. The log odds

ratio of interest is then the regression coefficient relating to the variable representing treatment group. In this model, it is again the case that results from larger trials are given greater weight in estimating the overall treatment effect than those from smaller trials, but the weighting is implicit in the model fitting procedure. In fact, logistic regression is asymptotically equivalent to the 'inverse-variance' method for log odds ratios. Logistic regression becomes advantageous, however, when it is necessary to adjust for additional covariates while still looking at an overall treatment effect. The confidence interval can easily be obtained using the variance of the estimate of the overall log odds ratio.

1.5.5 Example

The diuretics trials meta-analysis (Section 1.3.1) is used as an example data set to compare the four methods of estimation described in the previous sections. Furthermore, the 'general' and 'directional' tests and the Mantel-Haenszel test described in Section 1.4 are also used to analyse these data and the results of the tests compared.

Considering the issue of testing first, the odds ratio $a_i d_i / b_i c_i$ or the relative risk (RR) $a_i n_{i2} / c_i n_{i1}$ are both possible measurements of treatment effect. Both are considered here and hence the null hypothesis is that the odds ratio (or relative risk) is 1 in each study. Equivalently, since $\hat{\theta}_i$ is taken as the log odds ratio in practice, this is a test of the null hypothesis that the log odds ratio (or log relative risk) is zero in each study. It also follows that the weights $w_i = 1/var(\hat{\theta}_i)$ are taken as the reciprocal of the variance of the log odds ratio, $var(log\hat{O}R) = (1/a_i) + (1/b_i) + (1/c_i) + (1/d_i)$, or the reciprocal of the variance of the log relative risk, $var(log\hat{R}R) = (b_i/a_i n_{i1}) + (d_i/c_i n_{i2})$.

Since no assumptions are made, in any test, regarding the distribution of the different treatment effects under the alternative hypothesis H_1, the tests are always valid for the null hypothesis $H_0 : \theta_1 = ... = \theta_k = 0$. However, they do not test the null

Table 3: Test results for the diuretics trials data for the null hypothesis that there is no treatment effect in any of the trials

Test	Outcome measure	Test statistic observed	Degrees of freedom for χ^2	p-value
General	logOR	47.11	9	<0.001
General	logRR	45.90	9	<0.001
Specific	logOR	19.85	1	<0.001
Specific	logRR	17.28	1	<0.001
Mantel-Haenszel		21.63	1	<0.001

OR=odds ratio

RR=relative risk

hypothesis that the overall treatment effect is zero. In order to test $H_0 : \theta = 0$, the assumption of homogeneity of the treatment effects must hold, and only then may the alternative hypothesis be defined as $H_1 : \theta \neq 0$.

All tests carried out gave highly significant test statistics (Table 3), thus providing evidence against the null hypothesis. The tests using relative risk instead of odds ratio produced the same conclusions with the test statistic in each case being slightly smaller. The results indicate that there is strong evidence in the diuretics trials that at least one trial has a log odds ratio which is different from zero. Hence, the conclusions to be drawn regarding a meta-analysis from these tests are rather limited and so it can be seen why it is usually desirable to produce an estimate of an overall treatment effect.

The inverse-variance estimate, Mantel-Haenszel estimate and the Peto estimate are easily calculated, but computer software is required to carry out logistic regression. Hence, the statistical modelling package GLIM was used in order to obtain the logistic regression results [58]. For each of the 18 (9 trials x 2 treatment groups)

subgroups, the number of events of pre-eclampsia (r) was entered into the program, together with the number of patients (n). The dependent variable was therefore r, the error structure binomial with denominator n, and the link was logit. Factor variables were created for trial (levels 1−9) and for treatment group (1=control, 2=treatment). An additive model was then fitted including both trial and treatment group variables so that the estimate of overall treatment effect was obtained.

Table 4: Estimates of the overall odds ratio and its confidence interval for the diuretics trials data from the four different fixed effect methods

Method	Estimate of overall odds ratio	95% C.I. for odds ratio
Inverse-variance	0.67	(0.56,0.80)
Mantel-Haenszel	0.67	(0.56,0.80)
Peto method	0.66	(0.56,0.79)
Logistic regression	0.66	(0.56,0.79)

The results from all four methods produce almost exactly the same estimate of the overall odds ratio and almost the same 95% confidence interval (Table 4). This will not be the case for every set of data, however, since some estimators perform better than others under specific conditions. The inverse-variance method is generally to be preferred as it is asymptotically unbiased and consistent for all values of θ [59]. However, this method has a disadvantage in that it cannot be used if at least one of the studies in the meta-analysis has an event rate of zero. Furthermore, if the sample sizes are small, then the asymptotic normality assumptions will not hold [60, 61]. The same is true for small sample sizes with the Peto estimation method, where the validity of the confidence interval relies on the approximate normality of the score statistic under the null hypothesis of no overall treatment effect [40]. It has also been observed that the Peto estimate can yield extremely biased results when applied to sets of data where there is a large imbalance between the number of patients in the

treatment group and the number in the control group and also when the odds ratio is far from unity [55]. The Mantel-Haenszel estimate, in contrast, is robust when there are small frequencies and it can cope easily with zero cells [50, 62]. It does become unreliable, however, like the Peto method, when the true odds ratio is a long way from unity, but also when there is severe heterogeneity, although it is reasonably robust to moderate heterogeneity [62]. The confidence limits proposed for the Mantel-Haenszel estimator are also approximations, but have been found to perform well, even when the counts in individual strata are very small, provided the method does not break down with too many zeros [61, 63]. A better alternative, when frequencies are small, may be to use exact methods (Section 2.5) which are based on exact distribution theory [61]. It should also be noted that the choice of method will be limited when the outcome of interest is other than the odds ratio.

Although the consistency of the results in Table 4 might appear to suggest that the estimate obtained is a reliable indication of the true treatment effect, the choice between the different fixed effect estimates is not the real issue here. The problem is whether a fixed effect model is appropriate for this set of data at all, or whether the assumption of homogeneity is unrealistic.

1.6 Heterogeneity Across Studies

For the assumption of homogeneity to be valid, the assumption that θ_i is equal to θ for every study i (i=1,...,k) must be satisfied. If this condition does not hold then heterogeneity is present. Due to the fact that in the majority of meta-analyses there will be differences in study protocol, type of patient, and treatment duration and regimen between the trials, it would not be surprising if each study were to have a different underlying treatment effect. Hence, in practice, this clinical heterogeneity is likely to lead to statistical heterogeneity and therefore the breaking of the assumption underlying the fixed effect model.

A formal test of heterogeneity [38] may be performed using a statistic which is derived by considering the squared deviation of each study estimate from the true overall mean, $(\hat{\theta}_i - \theta)^2$, and then standardising by dividing by the study variance $v_i=1/w_i$. Then assuming that the w_i are fixed, $w_i(\hat{\theta}_i - \theta)^2$ has a χ^2_1 distribution and hence summing over all k studies produces a statistic $\sum_{i=1}^{k} w_i(\hat{\theta}_i - \theta)^2$ which has a χ^2_k distribution under the null hypothesis of homogeneity. However, the true value θ is, of course, unknown and must be replaced by the weighted mean estimate $\hat{\theta}_f$ (6). This means that 1 degree of freedom is lost from the null distribution and the actual test statistic used is given by

$$Q = \sum_{i=1}^{k} w_i(\hat{\theta}_i - \hat{\theta}_f)^2 \tag{12}$$

where $\hat{\theta}_f=\sum_{i=1}^{k} w_i\hat{\theta}_i/\sum_{i=1}^{k} w_i$, which is then compared to a χ^2_{k-1} distribution. It is assumed in the calculation of Q that the weights are known, when in practice they are estimated, and this issue is addressed in Chapter 5.

A non-significant result for the test of heterogeneity does not prove homogeneity, particularly since the test is not very powerful [35]. (The issue of power of the test will be investigated further in Chapter 4.) Therefore, even when a non-significant result is obtained, interpretation should still bear in mind the possibility of heterogeneity across the studies. In the example of the diuretics trials, however, it is clear that heterogeneity exists, since calculation of the test statistic given in (12) produces the value Q=27.27, which is highly significant ($p < 0.001$) when compared to the χ^2_8 distribution.

Peto presents a "natural approximate chi-square test" of the homogeneity assumption. The test is based on the use of the efficient statistic Z_i and Fisher's information V_i as defined in Section 1.5.3, substituted into the formula for Q (12) [4]. The statistic becomes simply the overall χ^2_k test minus the χ^2_1 specific test and is

given by

$$Q_P = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{V_i} - \frac{[\sum_{i=1}^{k}(O_i - E_i)]^2}{\sum_{i=1}^{k} V_i} \tag{13}$$

which has an approximate χ^2 distribution with $k^* - 1$ degrees of freedom, where k^* is the number of non-zero variances (k^* is usually equal to k). In the diuretics trials example, the results for this test agree with Q, with Q_P being equal to 29.3 (p <0.001).

The results of the logistic regression, obtained from the fit of the model in GLIM, also yield a test of heterogeneity by way of the deviance. The deviance is a measure of the fit of the model, and so if a significant 'trial by group' interaction (or equivalently heterogeneity) exists, there will be a lack of fit of the model, indicated by a large deviance in comparison to a χ^2 distribution. Again in the diuretics trials example, heterogeneity was found to be present since the deviance (29.4) indicated a significant lack of fit ($p < 0.001$). However, when small frequencies are present in the data, caution in the interpretation of the deviance is required, since the approximation to a χ^2 distribution may then be poor [58].

Thus, the results obtained from the fixed effect methods must be interpreted cautiously in the case of the diuretics trials data, since the homogeneity assumption on which they are founded is obviously not valid. Hence, the fixed effect estimates are of little value by themselves and in such circumstances should certainly not be presented without any reference to the heterogeneity which is present.

If heterogeneity is found to be present in a meta-analysis, which is common in medical research situations, then there are two options available. The first, which is generally to be preferred, is to look at the reasons behind the heterogeneity. This may involve consideration of the various characteristics of the original studies and an investigation of trial and patient differences [55, 64, 65]. It may thus be possible to

explain and, therefore, effectively to eliminate heterogeneity from the meta-analysis by identifying homogeneous subgroups of trials. However, it should be noted that such procedures are always post-hoc and must therefore be carried out with care [12]. If not all the variation can be explained in this way, or if an investigation is impossible due to the required data being unavailable, then an alternative approach is to incorporate the heterogeneity into the meta-analysis model by, for example, using the so called random effects model presented in the next section.

1.7 The Random Effects Method of Meta-Analysis

A standard random effects method based on the calculation of a moment estimator of the between-study variance is described in Section 1.7.1. An example is presented in Section 1.7.2 where the random effects results are compared with the fixed effect results, while Section 1.7.3 contains a discussion.

1.7.1 Standard random effects method

In situations where there is heterogeneity present, the random effects method for meta-analysis provides a way of incorporating between-study variability into the overall estimate. However, assumptions different from that of homogeneity must be made instead, and these are that the true treatment effect of each individual trial θ_i, $i = 1, ..., k$, is distributed with mean θ and a between-study variance σ_B^2. This implies that trials included in the meta-analysis are a random sample from an overall population of all such trials. Then each separate trial estimate $\hat{\theta}_i$, $i = 1, ..., k$, is assumed to have a distribution with mean θ_i and variance v_i. By setting up this model, an estimate of the between-study variance may be obtained, which can then be used in the calculation of a random effects estimate of the overall treatment effect θ.

One formula used to calculate $\hat{\sigma}_B^2$ [38], similar to that used in the random

effects one-way analysis of variance, is based on the Q statistic of heterogeneity (Section 1.6) and is derived by consideration of the expectation of Q (12). Rewriting Q as $\sum_{i=1}^{k} w_i(\hat{\theta}_i - \theta)^2 - (\sum_{i=1}^{k} w_i)(\hat{\theta} - \theta)^2$ in order that the expectation may be easily obtained gives

$$E(Q) = \sum_{i=1}^{k} w_i var(\hat{\theta}_i) - (\sum_{i=1}^{k} w_i) var(\hat{\theta}) \tag{14}$$

$$= (k-1) + \sigma_B^2 \left(\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i} \right) \tag{15}$$

Then using the method of moments, equating Q with the expected value of Q (15), and rearranging the resulting equation, an estimate of the between-study variance is obtained:

$$\hat{\sigma}_B^2 = \frac{Q - (k-1)}{\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}} \tag{16}$$

However, when Q is less than $(k-1)$, the estimate of σ_B^2 will be less than zero. Since a variance cannot take a negative value, the actual estimate of σ_B^2 used in practice is

$$max\{0, \hat{\sigma}_B^2\} \tag{17}$$

where $\hat{\sigma}_B^2$ is given in (16).

The unbiased random effects estimate of θ, $\hat{\theta}_r$ say, can then be found by calculating a weighted average of the individual estimates, as in formula (6), but with new weights w_i^*. These weights incorporate the additional component of variance and are given by $1/(v_i + \sigma_B^2)$. Hence, once more assuming that the weights, and therefore

σ_B^2, are known (the issue of allowing for the estimation of σ_B^2 is considered in Chapter 2), the random effects estimate of the overall treatment effect is given by

$$\hat{\theta}_r = \frac{\sum_{i=1}^{k} w_i^* \hat{\theta}}{\sum_{i=1}^{k} w_i^*} \tag{18}$$

To obtain confidence intervals it is necessary to make further assumptions about the form of the distribution of the treatment effects and the random effects, that is $\hat{\theta}_i \sim N(\theta_i, v_i)$ and $\theta_i \sim N(\theta, \sigma_B^2)$. (Chapter 3 addresses the problem of checking these distributional assumptions). The approximate variance of $\hat{\theta}_r$ is then given by

$$var(\hat{\theta}_r) = \frac{1}{\sum_{i=1}^{k} w_i^*}, \tag{19}$$

and hence confidence intervals for the log odds ratio may be obtained.

1.7.2 Example

The random effects estimate was calculated for the meta-analysis of the nine diuretics trials and the results were compared with those from the fixed effect methods (Section 1.5.5).

The random effects estimate of the overall odds ratio is smaller than the estimates obtained using any of the fixed effect methods (Table 4). Table 5 compares the results from the random effects method with those from the fixed effect inverse-variance method, which is in fact the random effects model where $\hat{\sigma}_B^2$ is taken equal to zero. It can also be seen from the comparison of results on the odds ratio scale that the 95% confidence interval for the random effects estimate is substantially wider than that for the fixed effect estimate.

The estimate of 0.230 for the between-study variance indicates that the vari-

Table 5: Comparison of the estimates of the overall treatment effect and its confidence interval from the inverse-variance fixed effect and random effects methods for the diuretics trials data

Method	Estimate of between-study variance $\hat{\sigma}_B^2$	Estimate of overall OR $(e^{\hat{\theta}})$	95% C.I. for e^{θ}
Fixed (inverse-variance)	0	0.67	(0.56,0.80)
Random	0.230	0.60	(0.40,0.89)

$\hat{\theta}=\hat{\theta}_f$ or $\hat{\theta}_r$

ability between studies is large in comparison to the variation within individual studies. There are only two of the nine trials (trials 4 and 8) which have an individual variance v_i greater than this between-study variance.

1.7.3 Discussion

The wider confidence interval associated with the random effects estimate is due to the extra variability introduced into the model by the between-study variance. In order to see why the random effects estimate is lower than the fixed effect estimate for these data, it is necessary to consider the allocation of weight. The simplest way to achieve this is to look at the percentage weight, defined as $(w_i/\sum_{j=1}^{k} w_j)\times 100\%$ for trial i, given to each of the individual estimates under the two models (Table 6). Including a between-study component of variance in the model has the effect of levelling out the weights. In the fixed effect method, where $\sigma_B^2=0$, the trial with the smallest variance, that is the most precise study, is given over half of the weight (54.6%). The next largest allocation, which is to trial 9, is much less at only 11.8%. Trial 8, in which only a very small number of events occurred, is given hardly any weight (1.2%).

These weights clearly contrast with those obtained from the random effects method. The additional variation means that, although trial 6 still receives the largest amount of weight, its share has been considerably reduced from 54.6% to 17.0%. The next largest weight is now not much less, being 13.9%, and all the other trials have been given the extra weight which has been taken away from trial 6. This means that the overall estimate from the random effects method is not so dominated by just one single trial, as it is in the fixed effect case. The fact that trial 6 has a odds ratio higher than the overall fixed effect estimate and that it loses weight in the random effects method pulls the overall estimate down. Also, this weight is redistributed primarily between trials 1, 2, 3, 5, and 7, three of which have odds ratios much lower than the fixed effect estimate, and this further explains why the estimate is lowered.

Table 6: Comparison of the percentage weights allocated to each of the diuretics trials in the fixed effect and the random effects methods

Trial	Percentage of total weight $(w_i / \sum_{j=1}^{k} w_j) \times 100$	
	Fixed effect	Random effects
1	5.0	10.7
2	6.8	11.9
3	4.5	10.2
4	2.7	7.9
5	7.0	12.0
6	54.6	17.0
7	6.6	11.8
8	1.2	4.5
9	11.8	13.9

w_i=weight allocated to study i, $i = 1, ..., k$

The random effects estimate, like the fixed effect estimate, should not simply be quoted without any questions regarding the validity of the model. The between-study variance has been estimated and should, therefore, ideally have its own measure of precision. The fact that in this example, as in many meta-analyses, σ_B^2 has only been estimated from a small number of trials means that the estimate will not be very precise. Furthermore, in calculating the confidence interval for θ, it has been assumed that the value of the between-study variance is known rather than estimated. This means that although the random effects confidence intervals, based on taking the variance as $1/\sum_{i=1}^{k} w_i^*$ under the assumption of normality, are wider than the fixed effect intervals, they are still likely to be too narrow, since they do not take into account the variability in $\hat{\sigma}_B^2$. These points are pursued further in Chapter 2.

A further point which has already been briefly mentioned, which relates to both the random effects and the fixed effect methods, is that they each assume that the individual study variances v_i are known, whereas in reality they must be estimated from the data. This problem was identified by DerSimonian and Laird [38] as an aspect of their method which required further investigation and they suggested that it may be preferable to use alternative estimators of the v_i to that proposed. The problem of estimating weights is considered later in the thesis (Chapter 5).

1.8 Displays in Meta-Analysis

The presentation of a statistical analysis is generally enhanced by graphical displays of the data. Meta-analysis is no exception in this respect, as a clear idea of all the trial estimates in relation to each other is required. Displays can also be useful when interpreting the amount of heterogeneity contained in a particular set of data. The amount of heterogeneity in any given meta-analysis can be tested using the statistic Q (Section 1.6). An estimate of the between-study variance can also be obtained which is an indication of the amount of heterogeneity present (Section 1.7.1). However,

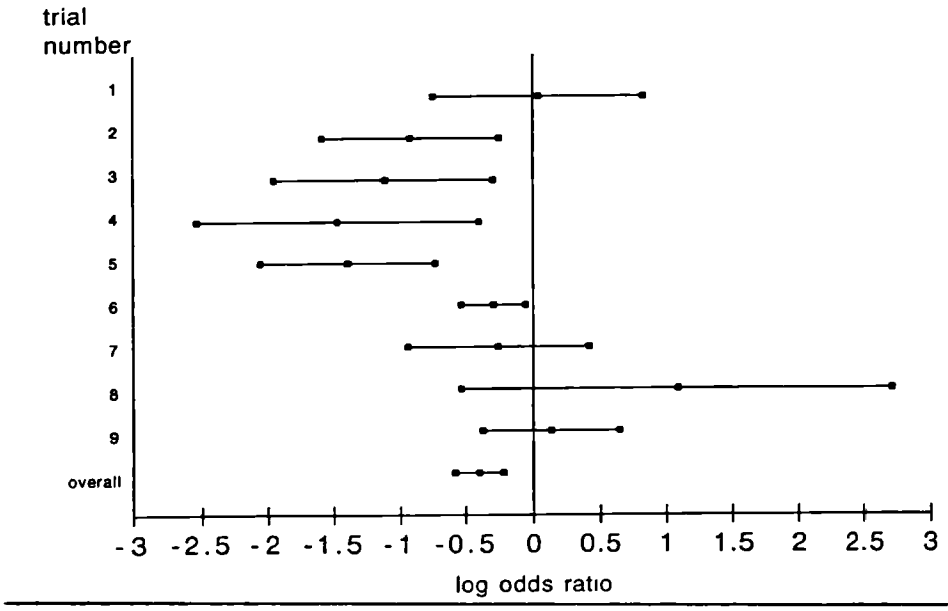
what the value of the between-study variance means in terms of heterogeneity can be difficult to comprehend. This section reviews ways in which graphical displays and plots can be used when presenting a meta-analysis.

The standard way to represent a meta-analysis data set in a diagrammatic way is to present each individual trial estimate of treatment effect $\hat{\theta}_i$, together with its 95% confidence interval on a single plot [33]. The overall fixed effect estimate $\hat{\theta}_f$ and its 95% confidence interval are usually plotted as well (Figure 1). From such a plot it can be seen whether there is much variation in the individual trial estimates $\hat{\theta}_i$, and, by looking at the overlap of the confidence intervals, how compatible they are with each other. Hence, some grasp of the amount of heterogeneity present in the data can be gained from these simple displays. These diagrams also provide information as to which trials produce a statistically significant treatment effect and which do not. Furthermore, it may be seen that the confidence interval for the overall estimate of treatment effect is much narrower than those of the individual trial estimates (Figure 1), indicating an increase in precision.

When odds ratios are the outcome measurement of interest, then it has been suggested by Galbraith [66] that, for two reasons, a display on the log scale is to be preferred to that on the linear scale. Firstly, the confidence intervals will be symmetrical, rather than asymmetrical as on the linear scale, and secondly, a unit change in the log odds ratio corresponds to a multiplication of the odds ratio by the same factor at any point on the scale. This means that 0.5 and 2.0, for example, are equidistant from 1. Also, the asymmetrical intervals on the linear scale may not always fit onto a conveniently scaled diagram if the precision of the $\hat{\theta}_i$ varies considerably. Trials are often ordered chronologically [46] on these diagrams, but may also be ordered by quality [36] or grouped such that trials with similar designs or characteristics are displayed together [67, 68].

However, regardless of the outcome measure or scale used, a disadvantage of

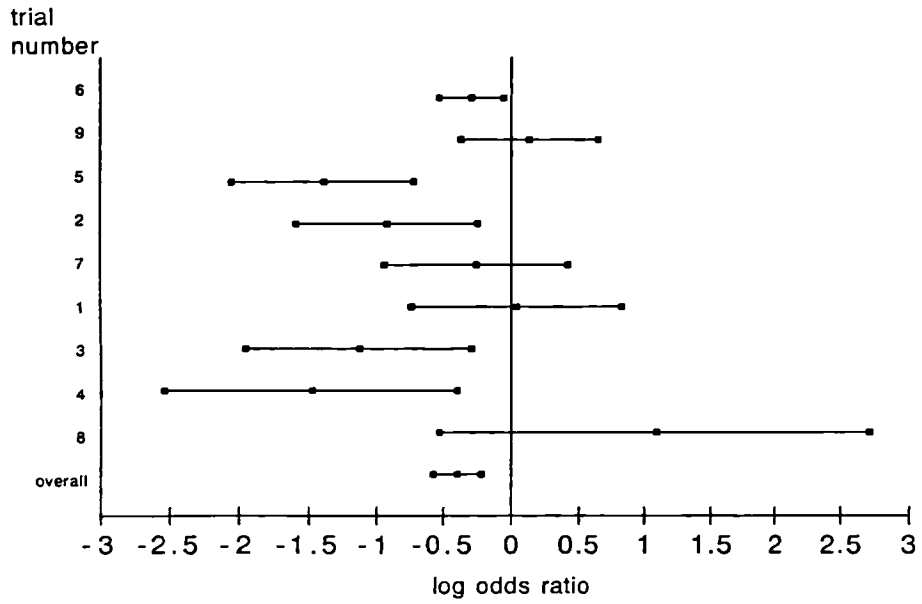
Figure 1: Standard meta-analysis diagram showing each ind vidual trial estimate of treatment effect together with its 95% C.I. for the diuretics trials data



this type of display has been identified. Trials which have more observations or events, thus providing more reliable information to the meta-analysis, have the narrowest confidence intervals. The visual impression of the plots can therefore be misleading in that the trials which are least informative tend to dominate the diagram as they have the widest confidence intervals. In order to get over this problem, a simple improvement to the diagram may be to order the trials with the most informative at the top and the least informative at the bottom, that is in increasing width of confidence interval (Figure 2).

Alternatively, the idea of representing the percentage weight allocated to each trial in the fixed effect model (or random effects model) by means of squares drawn on the individual confidence interval has also been used [64, 69, 70, 71] (Figure 3). The squares are such that their areas are proportional to the weight and so the larger the square, the more informative the trial. It may clearly be seen that in the

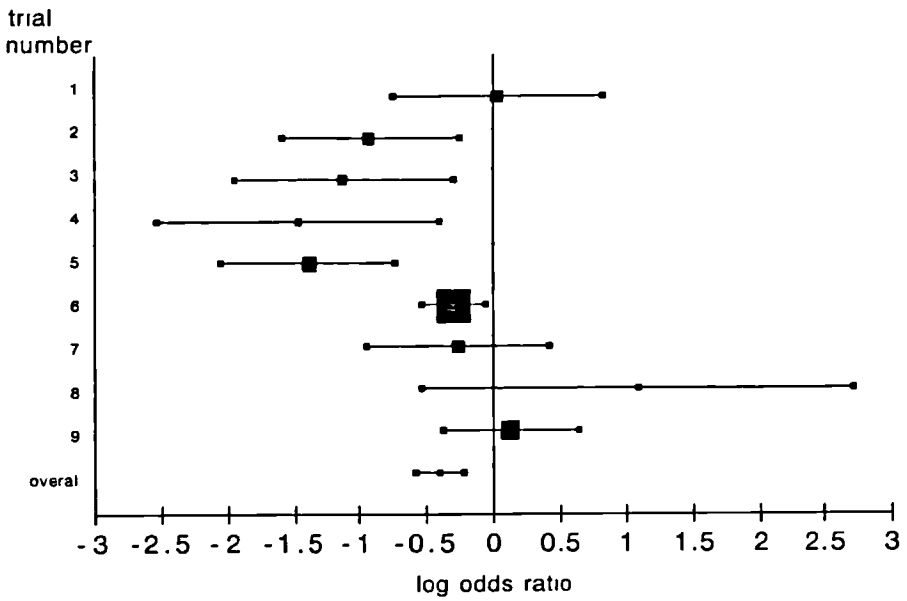
Figure 2: Standard meta-analysis diagram for the diuretics trials data with trials ranked from the most to the least informative



diuretics trials meta-analysis, a single trial takes most of the weight in the fixed effect model (Figure 3). The importance of looking at percentage weights has already been discussed (Section 1.7). Ways of incorporating this information into the displays is therefore to be encouraged.

A further adaption of the standard meta-analysis diagram suggested by Lau et al. [23] is to display a 'cumulative' meta-analysis, whereby a new overall estimate and new confidence interval is plotted as each new trial result is added in chronological order. This may be done using a fixed effect (Figure 4) or a random effects method (Figure 5). Typically the picture obtained will be of a series of increasingly narrow confidence intervals centring around an increasingly stable point estimate [36]. These plots, therefore, provide a continuous picture of the state of knowledge over the period in which the full set of trials were being carried out. It can clearly be seen at which stage the large trial (number 6) was incorporated into the analysis as there is a

Figure 3: Standard meta analysis diagram where the squares have areas proportional to the amount of information contributed to the fixed effect estimate



substantial decrease in the confidence interval on the fixed effect plot (Figure 4) at this point. The fact, however, that implicit multiple significance tests are being carried out here, that is one for each study added, means that adjustment of the p-values may be required, or alternatively that the display should not be interpreted formally.

Although useful as an initial look at the data to be included in a meta-analysis, these standard meta-analysis diagrams are not particularly informative for the purpose of investigating heterogeneity [27]. A diagram proposed by Galbraith [66] is an improvement in this respect and is discussed in Section 3.4.2.

Figure 4: Cumulative fixed effect meta-analysis diagram for the diuretics trials data

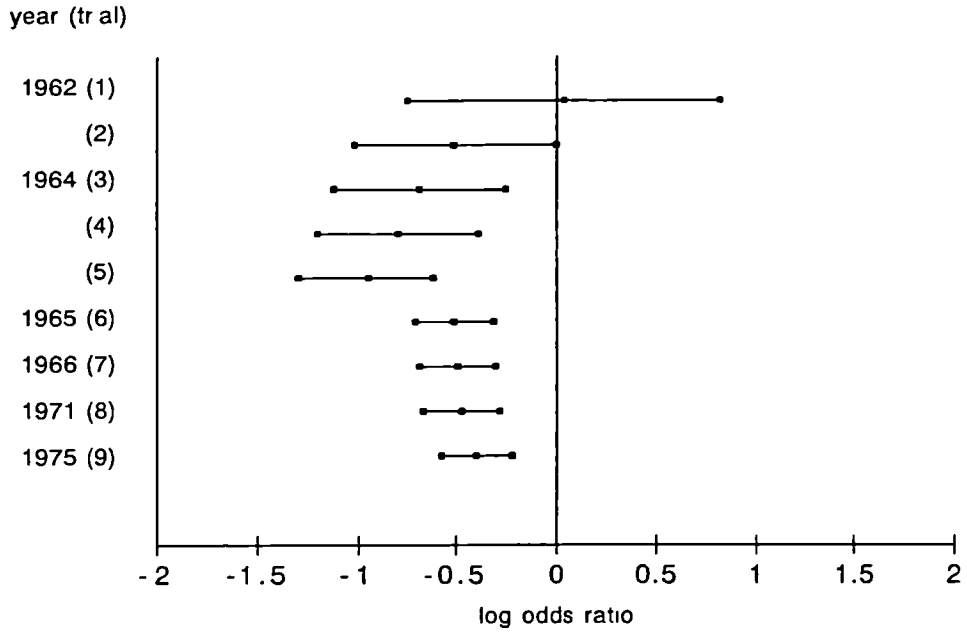
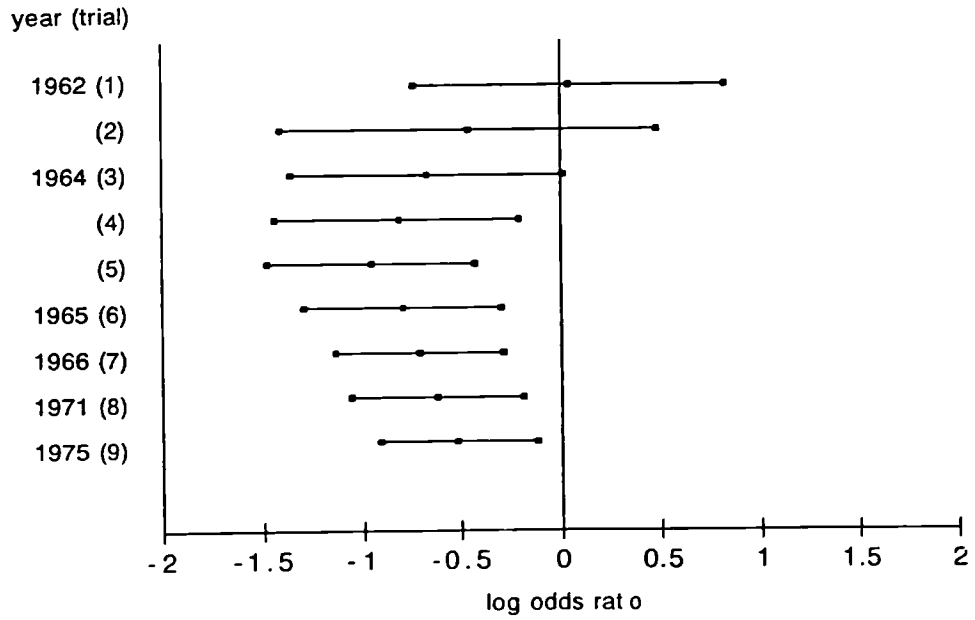


Figure 5: Cumulative random effects meta-analysis diagram for the diuretics trials data



1.9 Comparison of the Fixed Effect and the Random Effects Methods of Meta-Analysis

To complete Chapter 1, the interpretation of the fixed effect and random effects methods are discussed and differing views compared.

In the case where a group of studies have very similar protocols and study populations and in the absence of any statistical heterogeneity, it is generally agreed that the fixed effect estimate of an overall treatment effect is a valid way of summarising the data [36]. However, there is less consensus as to the best approach when clinical and statistical heterogeneity is present in a meta-analysis. Furthermore, there continues to be disagreement as to the interpretation of the fixed effect results and the value and appropriateness of the random effects model. Many researchers have advocated that the best approach to a meta-analysis, when heterogeneity exists, is to carry a full investigation of the sources of heterogeneity [21, 27].

The issue of hypothesis testing (Section 1.4) is much less controversial than that of the estimation of the overall treatment effect, since fewer assumptions are required. The approach to meta-analysis advocated by Peto [5] uses the idea of hypothesis testing rather than estimation. He favours the $(O-E)$ methods (Section 1.4 and 1.5.3) because "one can get all the asymptotic efficiency of logistic regression,...while avoiding the assumption that the relative risk is the same in each trial" and therefore prefers to term this approach "assumption free". The overall log odds ratio obtained using Peto's method is described as the "typical" log odds ratio, suggesting that it is an average value of the treatment effect derived from a selection of true fixed effect values. Peto then suggests the use of three standard deviations away from zero as being the reference standard for evidence of a treatment difference [5]. Hence, as long as it is remembered that the null hypothesis being tested is merely that of each treatment effect being equal to zero, then the approach outlined above is valid.

However, Peto's approach to estimation is far less clear and no specific exposition of these methods has been provided in the literature.

Various researchers [35, 72] have compared the fixed effect and the random effects approaches. Berlin et al. [35], when looking at the results of 22 meta-analyses, claimed that both methods often yield similar results, certainly at a qualitative (statistically significant versus non-significant) level. However, the variance estimates of the overall treatment effect did differ, with the random effects analysis being slightly more conservative. Although in the examples considered few qualitative differences existed, it should be noted that this generalisation will not hold for all sets of meta-analysis data. If the random effects model is more conservative, there will be cases when the two methods will produce different results at a qualitative level too; that is the fixed effect approach will produce a significant result and the random effects method a non-significant one. Berlin et al. [35] conclude from the study that the choice of methods may, therefore, depend on other more philosophical, rather than statistical, considerations. The random effects model has been criticised as being "a wholly wrong approach" [73], because it is answering the wrong and irrelevant question of "what would happen if we chose another treatment at random from the universe of treatments that we could choose and another population at random from the universe of populations?" The fixed effect model, on the other hand, it is claimed is addressing the question of interest [73]. Others express more doubt in the fixed effect model in the presence of heterogeneity [33, 72, 74].

When estimating the overall treatment effect, all fixed effect methods are making the same assumption of underlying homogeneity of the treatment effects, since the weights only take account of within-study variation and ignore any between-study variation. Rather than the point estimate being incorrect, the main problem with the fixed effect methods when heterogeneity is present, is that the standard error is too small thus meaning that the confidence interval is artificially narrow [12, 35, 38, 45].

The results of a simulation study [75] considering the issue of between-centre variance in a multicentre trial can be directly applied to meta-analysis since the data is simulated in exactly the same way in both cases using the same model. These results showed that the between-centre (or between-study) variation leads to the confidence intervals being inappropriately narrow. Peto has suggested the use of 99% instead of 95% confidence intervals [73], implying that greater evidence is required against the null hypothesis in a meta-analysis than a single clinical trial. However, 99% and 95% confidence intervals have different meanings and both will be spuriously narrow if calculated from a fixed effect model. A 99% confidence interval cannot simply replace a 95% confidence interval and reflect extra variation.

It is generally accepted that a thorough investigation of heterogeneity is required in order to attempt to explain the variation in the treatment effects. In the opinion of Greenland and Salvan [55], the choice between the two approaches is entirely secondary to the examination of inter-study heterogeneity. Indeed, Greenland [21] is sceptical of the random effects model in that he holds the view that it can conceal the fact that the overall estimate is a poor summary of the data. Similarly, Jenicek [65] believes that the analysis of heterogeneity should not be sacrificed in order to obtain some 'average' value. Large meta-analyses by those who favour the fixed effect approach [64, 69, 70] have also included an investigation of heterogeneity in the form of separate subgroup analyses. However, residual heterogeneity may remain unexplained, even after such an investigation, possibly because of some unmeasured or unreported study characteristic [36]. In such situations the random effects model may be useful, but should not be viewed as a panacea for any situation in which heterogeneity is large [36]. However, the random effects model does gives more appropriate confidence intervals than the fixed effect model [35].

The random effects model is certainly far from ideal as a method for obtaining an overall estimate of treatment effect. Firstly, because of the widely criticised and

unrealistic assumption that the trials in a meta-analysis are a random sample from a broader universe of trials, particularly when a specific form, usually normality, must be attached to this distribution. Furthermore, since the random effects method gives greater weight to smaller studies, the results may be emphasising poor evidence at the expense of good [33]. The estimate of the between-study variance is usually imprecise, being estimated from only a few trials, and is susceptible to bias [76] and hence this bias could affect the overall estimates. The random effects analysis is, therefore, probably best thought of as a check on the robustness of the conclusions from the fixed effect method to the failure in the assumption of homogeneity [76]. This sensitivity analysis will be of particular value when the sources of heterogeneity are intangible. Peto holds the view that the random effects approach can lead to the over cautious interpretation of results leading to a treatment being withheld from patients who would find it beneficial. However, if there is much between-study variation, then there may be uncertainty as to whether the treatment is of benefit, with different trials showing different results. If the random effects analysis indicates that the conclusions of the fixed effect analysis are valid then there will be additional cause for confidence in these results. If, however, the conclusions are different then it is likely that further information is required and no firm clinical conclusions can be drawn as to the benefit of the treatment. Dickerson and Berlin [36] state that the choice of the research question is critical in meta-analysis and a fairly general clinical question is often preferable to a very specific one, due to the heterogeneous nature of studies.

2 Extensions to the Standard Meta-Analysis Methods

The estimation methods for meta-analyses described in Chapter 1 are all commonly used standard techniques. However, none are completely satisfactory, especially in the presence of significant clinical and statistical heterogeneity. Chapter 2 considers extensions to these basic techniques and in particular addresses problems related to the estimation of the between-study component of variance in a random effects model.

The fact, that in a random effects model, the between-study variance is assumed to be known in the calculation of $var(\hat{\theta}_r)$ when in practice it must be estimated from the data, means that the standard random effects confidence interval is still too narrow. The initial sections of the chapter address this problem. Section 2.1 shows how a graphical method can be used as a sensitivity analysis to check the robustness of the estimate of the overall treatment effect to changes in the between-study variance. Section 2.2 proposes a likelihood method which produces both a confidence interval for the between-study variance and a confidence interval for the overall treatment effect which takes account of the fact that the between-study variance is estimated. Three practical examples are considered in Section 2.3 in order to illustrate this new methodology and this section also includes a discussion of the use of the information matrix to obtain approximate results from the likelihood model. Section 2.4 considers an alternative likelihood approach proposed by van Houwelingen, Zwindermann and Stijnen [45], based on the full conditional likelihood for binary outcomes and Section 2.5 shows how this method may be particularly useful when dealing with meta-analyses which include trials which have small numbers of, or even zero, events. Section 2.6 considers a Bayesian approach to meta-analysis, reviewing both empirical Bayes and fully Bayesian methods. Section 2.7 compares a proposed alternative moment estimator of the between-study variance with the standard DerSimonian and

Laird moment estimator (Section 1.7.1) and finally Section 2.8 contains a concluding discussion for Chapter 2.

2.1 Sensitivity Analysis

In the standard random effects model (Section 1.7.1) it is assumed that the weights w_i^* are known and therefore, that the value of the between-study variance is known. In practice σ_B^2 must obviously be estimated from the data, but the method does not take this imprecision into account when estimating θ. The fact that σ_B^2 is estimated means that there is an interest, when carrying out a meta-analysis, in the robustness of the estimate of the overall treatment effect to changes in the value of the between-study variance. Such an investigation gives an idea of the effect that an imprecise estimate of σ_B^2 may have on the estimate of treatment effect.

Section 2.1.1 explains the methods and uses an example to explain how a sensitivity plot may be produced. Section 2.1.2 discusses the resulting plot and investigates the reasons for the observed shape.

2.1.1 Methods

A plot of the estimate of θ against the between-study variance σ_B^2 may be used as a form of sensitivity analysis to assess how the estimate of the overall treatment effect θ varies across values of σ_B^2. Thus, when σ_B^2=0, the estimate of the treatment effect is simply that obtained from the fixed effect model using the inverse-variance method (Section 1.5.1). As σ_B^2 increases, the distribution of the weight between the trials in the meta-analysis becomes increasingly even. As σ_B^2 tends to infinity, the weights tend to equality and the overall estimate, therefore, tends to a simple unweighted average.

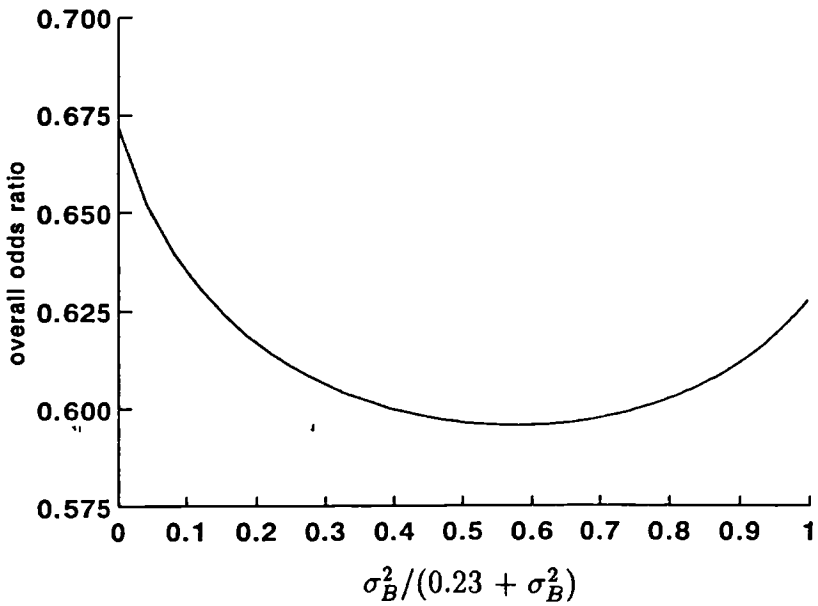
Using simple plots of $\hat{\theta}$ against σ_B^2 it is not possible to show the complete range of variation on a single graph. Hence, plots of the odds ratio against $\sigma_B^2/(\hat{\sigma}_B^2 + \sigma_B^2)$, where $\hat{\sigma}_B^2$ is the non-zero DerSimonian and Laird (D&L) moment estimate of the between-study variance (Section 1.7.1) were used. In the context of the meta-analysis of the diuretics trials (Section 1.3.1), the estimate of σ_B^2 is 0.23. The sensitivity plot of the estimated odds ratio, which is actually $e^{\hat{\theta}}$, is therefore plotted against $x = \sigma_B^2/(0.23 + \sigma_B^2)$ in Figure 6, so that $x = 0$ corresponds to σ_B^2=0, $x = 0.5$ to σ_B^2=0.23 (the D&L moment estimator) and x=1 to σ_B^2=∞, and the whole range of σ_B^2 is reduced to a finite scale. The choice of the moment estimate of σ_B^2 in the term $\sigma_B^2/(\hat{\sigma}_B^2 + \sigma_B^2)$ is rather arbitrary, as any constant number, such as 1 for example, could have been used and is, in fact, necessary in the situation where $\hat{\sigma}_B^2$ is zero. The use of $\hat{\sigma}_B^2$ does however mean that for $\hat{\sigma}_B^2$>0, the random effects estimate is always situated in the centre of the plot, that is at 0.5 on a scale of 0 to 1. This ensures that attention is always focused on the most important part of the plot, irrespective of the numerical value of the random effects estimate.

The diuretics trials data have an overall odds ratio (Figure 6) which decreases from the fixed effect estimate to a minimum (which happens to be close to the random effects estimate) and then increases again until it reaches the 'equal weighting' value. This type of pattern is not necessarily produced by these sensitivity plots, and the shape of a plot for any particular set of data is not easily predicted. The behaviour, particularly for large σ_B^2, tends to depend on the treatment effects in the smallest trials.

2.1.2 Discussion

In order to explain the shape of the plot which emerged for the diuretics trials data, the percentage weighting allocated to each trial was obtained for a range of values of σ_B^2. The main interest lies in the explanation of the minimum value and subsequent

Figure 6: Sensitivity plot showing how the overall odds ratio varies with the between-study variance (σ_B^2) for the diuretics trials meta-analysis



0.23=random effects estimate of between-study variance $\hat{\sigma}_B^2$

increase in overall odds ratio. Hence, weightings corresponding to values of σ_B^2 between 0 and 1 were studied since this range includes the area of the plot where the minimum occurs.

Trial number 8 (Table 7), which has a large individual odds ratio (OR=2.971) but is the most imprecise trial, gets allocated increasing weight as σ_B^2 increases and it is this that raises the overall estimate after the minimum value has been reached. Trial 8 has the lowest initial weighting and is consequently the last whose weight levels out. At a σ_B^2 value of approximately 0.5–0.6, the weightings for all the other trials are approximately equal (Table 7), and hence the major effect on the change in the overall odds ratio after this point must be that of trial 8 receiving increased weight.

Before the minimum is reached, the main influence which leads to a decrease in the estimate of the overall treatment effect is the rapid loss of weight of the largest trial (trial 6), which has an individual estimate higher than the overall fixed effect estimate. In the fixed effect estimate (σ_B^2=0), trial 6 is allotted over half of the total weight, while in the random effects estimate much of this weight has been redistributed to the other trials. Four (numbers 2, 3, 4 and 5) out of the other seven trials which receive this weight in the initial redistribution (i.e. all trials except trial 8) have low individual odds ratios (Table 7) and this would appear to explain the reduction in the overall estimate. Furthermore, it is interesting to note that the overall odds ratio begins to increase again soon after the random effects estimate and at a value of σ_B^2 which is far from extreme. Hence, the most imprecise trial would appear to have some influence on the random effects estimate of the overall treatment effect.

The sensitivity analysis was repeated with the small trial (trial 8) being excluded in order to see whether, indeed, the influence of this trial was important. This second analysis did produce a quite substantially different estimate of the overall odds ratio under the random effects model (Table 8). The estimate of the between-

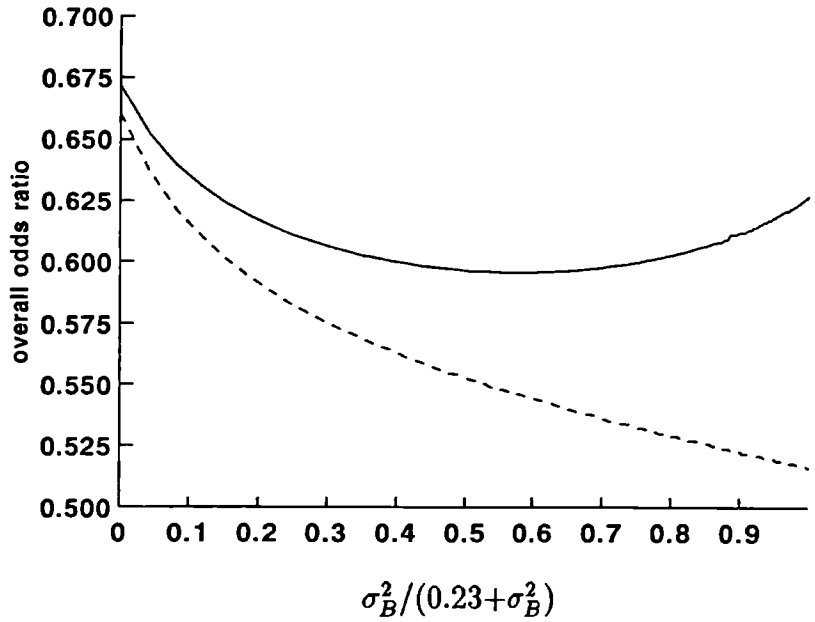
Table 7: Percentage weights allocated to each trial for different values of the between-study variance

Trial	Weight allocated to trial (%)											Odds ratio
	Between-study variance (σ_B^2) ($x = \sigma_B^2/(0.23+\sigma_B^2)$)											
	0.0 (0.0)	0.1 (0.30)	0.2 (0.47)	0.3 (0.57)	0.4 (0.64)	0.5 (0.69)	0.6 (0.72)	0.7 (0.75)	0.8 (0.78)	0.9 (0.80)	1.0 (0.81)	
1	5.0	9.7	10.5	10.9	11 0	11.1	11.1	11.2	11.2	11.2	11.2	1.04
2	6.8	11.6	11.9	12.0	11.9	11.7	11.8	11.8	11.7	11.7	11.6	0.40
3	4.5	9.1	10.0	10.4	10.7	10.8	10.9	11.0	11.0	11.0	11.0	0.33
4	2.7	6.3	7.6	8.3	8.8	9.2	9.4	9.6	9.8	9.9	10.0	0.23
5	7.0	11.8	12.1	12.1	12.0	11.9	11.9	11.8	11.8	11.7	11.7	0.25
6	**54.6**	**22.0**	**17.6**	**15.9**	**14.9**	**14.2**	**13.8**	**13.5**	**13.2**	**13.0**	**12.8**	**0.74**
7	6.6	11.4	11.8	11.9	11.9	11.8	11.8	11.7	11.8	11.6	11.6	0.77
8	**1.2**	**3.2**	**4.3**	**5.1**	**5.7**	**6.2**	**6.6**	**6.9**	**7.2**	**7.5**	**7.7**	**2.97**
9	11.8	15.0	14.1	13.6	13.2	12.9	12.7	12.5	12.4	12.3	12.2	1.15

study variance, however, remained almost the same, being only slightly smaller at 0.21. Consequently, the associated confidence interval for the overall odds ratio was marginally narrower and was shifted in position. The plot also differed in shape, no longer increasing after reaching a minimum value, but decreasing monotonically to the 'equal weighting' estimate (Figure 7). This was much lower than the 'equal weighting' estimate obtained for the full set of data.

For the full data set, the range of possible values of $e^{\hat{\theta}}$ is between 0.5956 and 0.6717 (Figure 6), and hence the variation in the point estimate is not particularly large. However, as σ_B^2 increases, the certainty with which the value of θ may be estimated decreases and hence the 95% confidence interval for θ gets wider as σ_B^2 increases. Confidence intervals may also be illustrated on the sensitivity plots. For each value of σ_B^2, the variance from the random effects model may be calculated using

Figure 7: Sensitivity plot showing how the overall odds ratio varies with the between-study variance (σ_B^2) comparing the full set of data with that excluding trial 8 for the diuretics trials meta-analysis



Key

— Full data

- - Excluding trial 8

0.23=random effects estimate of between-study variance $\hat{\sigma}_B^2$ for full data

0.21=random effects estimate of between-study variance $\hat{\sigma}_B^2$ excluding trial 8

Table 8: A comparison of the results for the full data with those from the data excluding trial 8

Case	Type of estimate	Estimate of between-study variance ($\hat{\sigma}_B^2$)	Estimate of overall odds ratio ($e^{\hat{\theta}}$)	95% C.I. for e^{θ}	Width of C.I.
Full data	Fixed	0	0.67	(0.56,0.80)	0.24
	Random	0.23	0.60	(0.40,0.89)	0.49
Without trial 8	Fixed	0	0.66	(0.55,0.79)	0.24
	Random	0.21	0.56	(0.37,0.82)	0.45

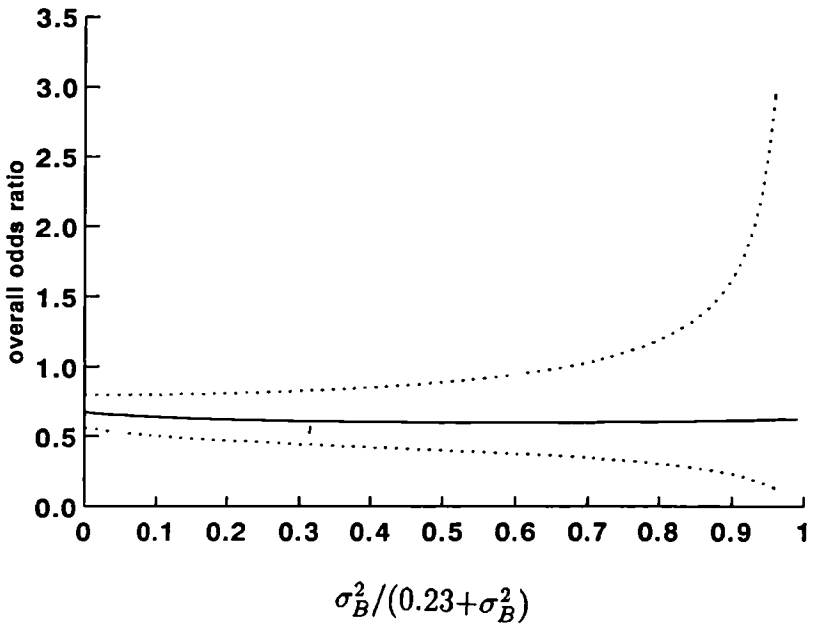
$var(\hat{\theta}_r) = 1/\sum_{i=1}^{k} w_i^*$ and so the corresponding confidence interval can be obtained. The plot (Figure 8) indicates how the initial increase in the width of the confidence interval is gradual, but how for extreme values of σ_B^2, the interval increases in width very rapidly, with the upper bound tending towards infinity.

The technique presented in this Section is useful for checking the robustness of any conclusions drawn from a fixed effect or a random effects meta-analysis. Since σ_B^2 cannot be estimated very precisely, especially when there is only a small number of trials in the analysis, the technique reveals whether this creates a problem in relation to the conclusion being drawn. If the sensitivity plot shows very little change over a range of different σ_B^2, then the greater the confidence in the conclusions. However, if the plot shows that the estimate and confidence intervals change markedly, then extra caution should be expressed in the results.

2.2 Maximum Likelihood Approach to Meta-Analysis Based on Marginal Distributions

A likelihood approach to meta-analysis is proposed in this Section, which is shown to offer certain improvements over the standard random effects method (Section 1.7.1).

Figure 8: Sensitivity plot showing how the overall odds ratio and its 95% confidence interval vary with the between-study variance (σ_B^2) for the diuretics trials meta-analysis



Key

— odds ratio

... 95% confidence limits

0.23=random effects estimate of between-study variance $\hat{\sigma}_B^2$

The model is introduced in Section 2.2.1, while Section 2.2.2 describes how confidence regions may be obtained and Section 2.2.3 derives confidence intervals for both θ and σ_B^2 from the relevant profile likelihoods.

2.2.1 Introduction

The standard random effects confidence interval for θ is too narrow as it makes no allowance for the imprecision in the estimate of σ_B^2. Furthermore, there is no published method for calculating a confidence interval for σ_B^2 itself. In meta-analyses, which commonly include only a small number of trials, the estimate of σ_B^2 will not be very precise and so any such confidence interval would be wide. Initially an analogy to the one-way analysis of variance was pursued in order to obtain a confidence interval for σ_B^2, but this approach could only be applied in certain situations and proved problematic for the general case. Both the problems mentioned above can, however, be solved by using a likelihood approach.

The random effects model was set up as described in Section 1.7.1 with the distributional assumptions of normality. Under this model, the marginal distribution of each individual estimated treatment effect $\hat{\theta}_i$, $i = 1, ..., k$, is, therefore, normal with mean θ and variance $(v_i + \sigma_B^2)$. Hence the contribution from study i to the likelihood for θ and σ_B^2 is,

$$L_i(\theta, \sigma_B^2) = \frac{1}{\sqrt{2\pi(v_i + \sigma_B^2)}} exp \left\{ \frac{-(\hat{\theta}_i - \theta)^2}{2(v_i + \sigma_B^2)} \right\} \qquad (20)$$

For a meta-analysis involving k independent studies, the full likelihood is the product of the individual study likelihoods. The log-likelihood is, however, simpler to work with, and is given by,

$$l(\theta, \sigma_B^2) = -\sum_{i=1}^{k} \frac{1}{2} ln 2\pi (v_i + \sigma_B^2) - \sum_{i=1}^{k} \frac{(\theta_i - \theta)^2}{2(v_i + \sigma_B^2)} \tag{21}$$

To obtain maximum likelihood estimates (MLEs) of θ and σ_B^2 that is $\hat{\theta}_l$ and $\hat{\sigma}_{Bl}^2$, the partial derivatives of equation (21), $\partial l/\partial\theta$ and $\partial l/\partial\sigma_B^2$, are set to zero and the resulting expressions rearranged [77]. The equations thus obtained, (22) and (23), are then solved in an iterative manner, beginning by substituting an initial value of $\hat{\sigma}_{Bl}^2$ in equation (23),

$$\hat{\theta}_l = \frac{\sum_{i=1}^{k} \frac{\hat{\theta}_i}{(v_i + \hat{\sigma}_{Bl}^2)}}{\sum_{i=1}^{k} \frac{1}{(v_i + \hat{\sigma}_{Bl}^2)}} \tag{22}$$

$$\hat{\sigma}_{Bl}^2 = \frac{\sum_{i=1}^{k} \frac{(\hat{\theta}_i - \hat{\theta}_l)^2 - v_i}{(v_i + \hat{\sigma}_{Bl}^2)^2}}{\sum_{i=1}^{k} \frac{1}{(v_i + \hat{\sigma}_{Bl}^2)^2}} \tag{23}$$

The actual form given in (23) is not the simplest expression possible for $\hat{\sigma}_{Bl}^2$ and is obtained by subtracting $\sum_{i=1}^{k} v_i/(v_i + \hat{\sigma}_{Bl}^2)^2$ from either side of the initial equation, but it is the most convenient form for the implementation of the iteration process. Alternatively, the MLEs may be obtained directly from the log-likelihood given in (21) using, for example, Splus [78]. This implements a routine which produces a range of points from the joint likelihood $l(\theta, \sigma_B^2)$ using a grid of values of θ and σ_B^2. The program then finds the maximum value of $l(\theta, \sigma_B^2)$ and hence $\hat{\theta}_l$ and $\hat{\sigma}_{Bl}^2$ also.

2.2.2 Confidence regions

The joint log-likelihood of θ and σ_B^2 in (21) can be calculated and three-dimensional plots obtained using Splus [78]. This same likelihood can also be displayed on a contour plot where the contours join all the points which have the same log-likelihood

and hence represent likelihood-based confidence regions. To obtain a joint confidence region for both parameters θ and σ_B^2 the fact that $-2\{l(\theta,\sigma_B^2) - l(\hat{\theta}_l,\hat{\sigma}_{Bl}^2)\}$ has an asymptotic χ_2^2 distribution [79] is used. The approximate 95% likelihood-based confidence region is thus given by all pairs of θ and σ_B^2 which satisfy,

$$l(\theta,\sigma_B^2) > l(\hat{\theta}_l,\hat{\sigma}_{Bl}^2) - 5.991/2 \tag{24}$$

where 5.991 is the 95% point of the χ_2^2 distribution.

2.2.3 Profile likelihoods

The profile log-likelihood can be used in order to find confidence intervals for each of θ and σ_B^2. The profile log-likelihood is the log-likelihood for one parameter, which takes into account the fact that the other parameter is unknown and must be estimated. Hence, the profile log-likelihood for σ_B^2 is obtained by replacing θ in equation (21) by the maximum likelihood estimate $\hat{\theta}_l$,

$$\hat{\theta}_l = \frac{\sum_{i=1}^{k} \frac{\hat{\theta}_i}{(v_i+\sigma_B^2)}}{\sum_{i=1}^{k} \frac{1}{(v_i+\sigma_B^2)}} \tag{25}$$

for each given value of σ_B^2, that is profile out θ. The profile log-likelihood, $l^*(\sigma_B^2) = l(\hat{\theta}_l(\sigma_B^2),\sigma_B^2)$, where $\hat{\theta}_l(\sigma_B^2)$ is the MLE of θ for a given σ_B^2, may then be plotted against σ_B^2. It requires more work to obtain the profile likelihood for θ, since the maximum likelihood estimate of σ_B^2 cannot be written in terms of θ alone. The maximum likelihood estimate of σ_B^2 for any given θ satisfies the equation

$$\hat{\sigma}^2_{Bl} = \frac{\sum_{i=1}^{k} \frac{(\hat{\theta}_i - \theta)^2 - v_i}{(v_i + \hat{\sigma}^2_{Bl})^2}}{\sum_{i=1}^{k} \frac{1}{(v_i + \hat{\sigma}^2_{Bl})^2}} \tag{26}$$

Hence, in this case, the profile log-likelihood has to be found numerically, whereby for each given θ equation (26) is solved to find $\hat{\sigma}^2_B$. This maximum likelihood estimate is then used to obtain the corresponding point of the log-likelihood, $l^*(\theta) = l(\theta, \hat{\sigma}^2_{Bl}(\theta))$, using (21).

In order to obtain confidence intervals from the profile log-likelihoods, the fact that $-2\{$difference in profile log-likelihoods$\}$ has an asymptotic χ^2 distribution is utilised. Hence, $-2\{l^*(\sigma^2_B) - l^*(\hat{\sigma}^2_{Bl})\}$ and $-2\{l^*(\theta) - l^*(\hat{\theta}_l)\}$ each have approximately a χ^2_1 distribution [79], the degrees of freedom being the difference between the number of unknown parameters. It then follows that the 95% confidence intervals are given by all values of the parameters which satisfy the equations

$$l^*(\sigma^2_B) > l^*(\hat{\sigma}^2_{Bl}) - 3.84/2 \tag{27}$$

and

$$l^*(\theta) > l^*(\hat{\theta}_l) - 3.84/2 \tag{28}$$

where 3.84 is the 95% point of the χ^2_1 distribution.

A test for heterogeneity may also be derived from these profile log-likelihoods, since a null hypothesis of homogeneity is equivalent to $H_0 : \sigma^2_B = 0$. The one-sided alternative hypothesis is then $H_1 : \sigma^2_B > 0$, under the assumption of a normally distributed random effects model. The relevant likelihood ratio statistic to test for heterogeneity is therefore $LRT = \sqrt{2\{l^*(\hat{\sigma}^2_{Bl}) - l^*(0)\}}$ which can be compared to the standard normal distribution to obtain a one-sided p-value.

2.3 Practical Considerations

Three contrasting examples will now be considered in Sections 2.3.1–2.3.3 to illustrate how taking into account the imprecision in the estimate of σ_B^2 affects the confidence interval for θ. Section 2.3.4 then considers use of the information matrix as a way of approximating the confidence intervals for θ and σ_B^2 and Section 2.3.5 is a discussion.

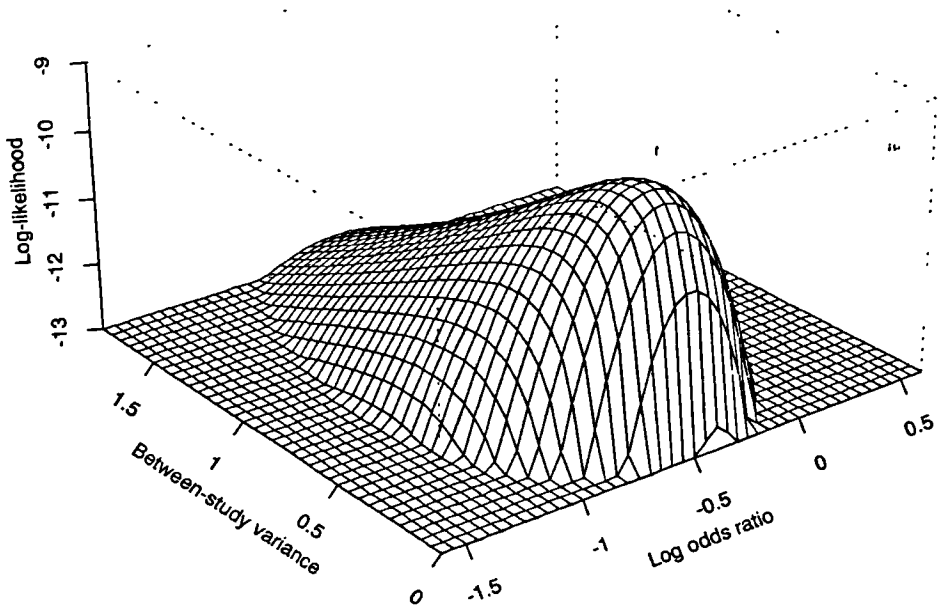
2.3.1 Example 1: Diuretics trials meta-analysis

For the diuretics trials meta-analysis (Section 1.3.1), it can be seen from the three-dimensional likelihood plot (Figure 9), as well as from the profile log-likelihood (Figure 10), that the likelihood of the between-study variance is, as expected, very asymmetric. The profile log-likelihood for the overall treatment effect indicates that the likelihood of θ is much more symmetric in shape (Figure 11). However, unlike the standard methods (Sections 1.5 and 1.7), using the profile likelihood does not force the confidence interval for either parameter to be symmetric.

The contour plot (Figure 12) is difficult to interpret and is best used to obtain an idea of the shape of the joint log-likelihood surface. It does, however, indicate that the possible ranges of σ_B^2 and θ are very much wider than the individual confidence intervals suggest. It can be seen that the 95% likelihood-based confidence region includes values of θ greater than 0 which indicates the possibility of no treatment effect.

The estimates and their corresponding confidence intervals based on the likelihood model can be compared to those obtained from both the fixed effect method and the standard random effects method using the D&L moment estimator of σ_B^2 (Table 9). The maximum likelihood estimates agree well with the standard random effects estimates. The fixed effect estimate of the overall treatment effect is slightly

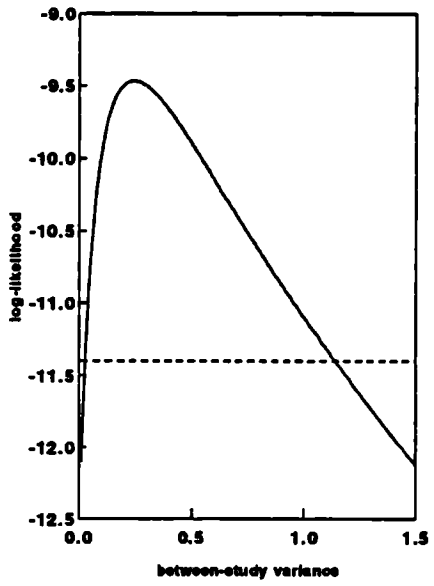
Figure 9: Bivariate distribution of the overall log odds ratio and the between-study variance for the diuretics trials meta-analysis



larger than the two random effects estimates. However, the fact that a reasonably large estimate of the between-study variance is obtained indicates a lack of homogeneity in this set of studies. The likelihood ratio test for heterogeneity produces a highly significant result (LRT=2.53, $N(0,1)$ p=0.006). However, the Q statistics for heterogeneity (Section 1.6) is even more significant (Q=27.27, χ^2_8 p=0.0007) and therefore appears to be more powerful than the likelihood ratio test in this example.

The confidence interval for the between-study variance is wide, reflecting the fact that only nine studies are included in this meta-analysis, meaning that the between-study variance is imprecisely estimated. Allowing for the estimation of the between-study variance means that the likelihood based confidence interval for the overall treatment effect is wider than that obtained by the standard random effects method. Nevertheless, bearing in mind the large imprecision in the estimate of σ^2_B, the increase in the width of the confidence interval for e^{θ} is relatively small.

Figure 10: Profile likelihood for the between-study variance for the diuretics trials meta-analysis



Key

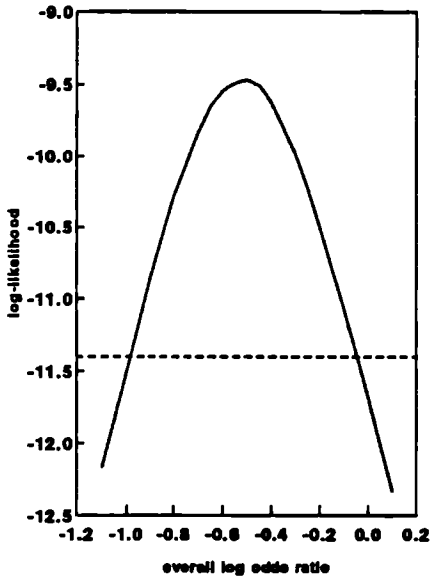
- - maximum log-likelihood − 1.92

0.24=MLE of between-study variance

Table 9: Comparison of the results from three meta-analysis methods for the diuretics trials data

Method	Estimate of between-study variance ($\hat{\sigma}_B^2$)	95% C.I. for σ_B^2	Estimate of overall odds ratio ($e^{\hat{\theta}}$)	95% C.I. for e^{θ}
Fixed effect	0.00	-	0.67	(0.56,0.80)
Random effects				
Standard	0.23	-	0.60	(0.40,0.89)
Likelihood	0.24	(0.03,1.13)	0.60	(0.37,0.95)

Figure 11: Profile likelihood for the overall log odds ratio for the diuretics trials meta-analysis

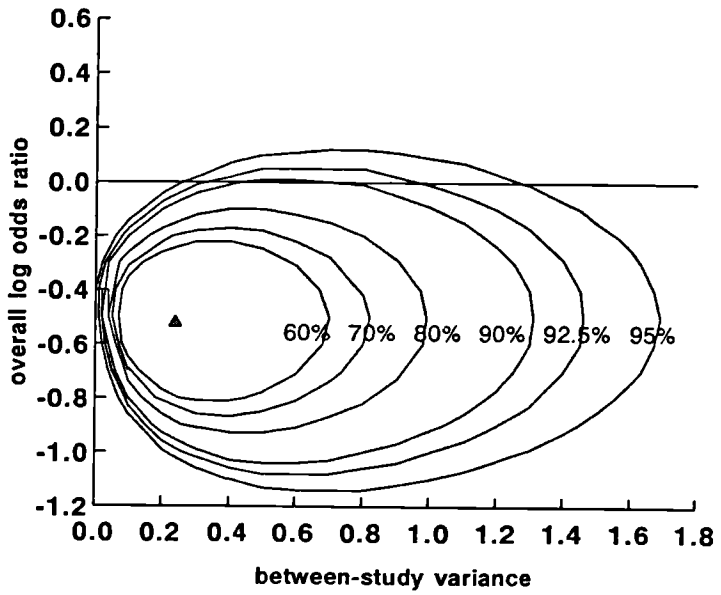


Key

- - maximum log-likelihood − 1.92

−0.51=MLE of overall log odds ratio

Figure 12: Contour plot for the bivariate distribution of the overall log odds ratio and the between-study variance for the diuretics trials meta-analysis



Key

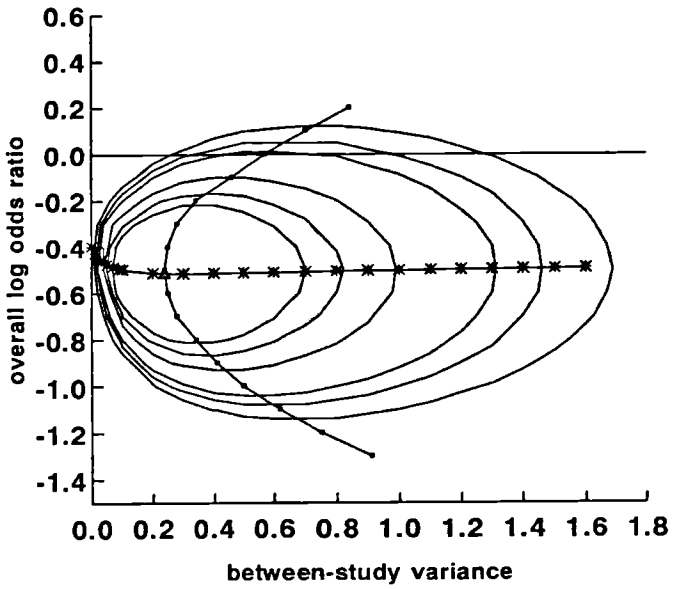
▲ maximum likelihood

The sensitivity plots described in Section 2.1 provide a guide to the effect that the imprecision in the estimate of σ_B^2 will have on $\hat{\theta}$ (or $\hat{\theta}_l$ as $\hat{\theta}_l$=$\hat{\theta}$ for a given σ_B^2) and hence how much the likelihood based confidence interval will differ from the standard random effects interval. The important characteristic of the sensitivity plots in this respect is what happens to the estimate of θ in the region of $\hat{\sigma}_{Bl}^2$. If the estimate of θ remains constant across the values of σ_B^2 contained in the 95% likelihood-based confidence interval, then the likelihood-based confidence interval for θ will be no different to that obtained from the standard random effects method. However, the greater the variation in $\hat{\theta}$ over this region of interest, the greater the increase in width of the interval for θ. The imprecision of the estimation of σ_B^2 does not have a direct influence on the width of the confidence interval for θ. However, the more imprecise the estimate of σ_B^2, the larger the range of influential values and hence the greater the chance of a variation in $\hat{\theta}$ in the range of interest.

Figure 6 shows that the estimate of the overall odds ratio $e^{\hat{\theta}}$ changes by only 0.03, approximately, in the region covered by the likelihood-based confidence interval for σ_B^2. This suggests that the estimate of the between-study variance does not have much influence on the overall estimate in this example, and hence a large increase in the width of the confidence interval for θ when using the likelihood method compared to the standard random effects method would not be expected. The same information may also be gained by considering the plot of the change in the MLE of the 'nuisance' parameter for different values of the parameter of interest on a contour plot (Figure 13). The value of $\hat{\theta}_l$ (=$\hat{\theta}$) changes very little when looking at the profile likelihood of σ_B^2, and it is only for very small values of σ_B^2 where there is a marked difference. Hence, the profile log-likelihood is almost the same as a cross-section cut through the joint likelihood at the maximum likelihood value of θ. This again suggests that the estimate of the overall mean does not depend greatly on the value of σ_B^2. In contrast, the value of $\hat{\sigma}_{Bl}^2$ does change quite considerably over different values of θ (Figure 13). As θ increases the value of $\hat{\sigma}_{Bl}^2$ decreases to the maximum likelihood estimate and

then increases again. This pattern is to be expected since $\hat{\sigma}^2_{Bl}$ will obviously be larger as θ moves away from the maximum likelihood estimate.

Figure 13: Contour plot showing how the estimates of the log odds ratio and the between-study variance change



Key

▲ maximum likelihood

— likelihood estimate of between-study variance $\hat{\sigma}^2_{Bl}$

-*- likelihood estimate of overall log odds ratio $\hat{\theta}_l$

Contours as in Figure 12

In moving from the simple fixed effect model to the likelihood model, which allows both for the heterogeneity and the estimation of the between-study variance, the certainty with which conclusions may be drawn from this meta-analysis changes. Although the point estimate of the overall treatment effect alters little, the increased width of the related confidence interval reduces the certainty of the conclusions (Table 9). The fixed effect interval is narrow and corresponds to a highly significant treatment benefit. However, in the likelihood analysis where the interval has in-

creased substantially from the fixed effect analysis, the treatment benefit of diuretics is substantially less significant, with the upper limit of the 95% confidence interval being only slightly below unity.

2.3.2 Example 2: A multicentre trial

The second example uses data from the Medical Research Council's multicentre trial of the treatment of mild hypertension (Section 1.3.2). This allows an example with a large number of 'trials' (i.e. centres in this example) to be considered. Furthermore, since the outcome considered here is the reduction in diastolic blood pressure in mmHg between entry to the trial and one year after entry, the outcome measure is continuous, which also contrasts with the first example. The likelihood methodology, however, carries through in an exactly similar way as for the log odds ratio.

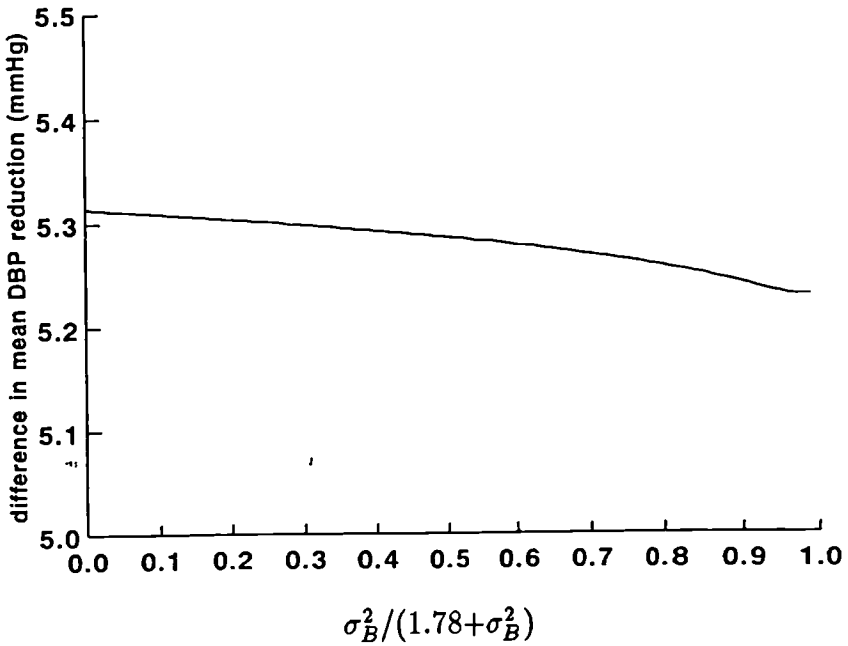
Table 10: Comparison of the results from three meta-analysis methods for the mild hypertension trial data

Method	Estimate of between-study variance ($\hat{\sigma}_B^2$)	95% C.I. for σ_B^2	Estimate of overall odds ratio ($e^{\hat{\theta}}$)	95% C.I. for e^{θ}
Fixed effect	0.00	-	5.31	(5.03,5.59)
Random effects				
Standard	1.81	-	5.29	(4.94,5.63)
Likelihood	1.78	(0.83,3.05)	5.29	(4.94,5.63)

In this example, there is considerable evidence of heterogeneity between centres (Q=278.12, χ^2_{188} p <0.0001), but there are no real differences between the standard random effects results and the likelihood results (Table 10). In contrast to the previous example, the likelihood ratio test (LRT=19.8, $N(0,1)$ p <0.00005) produces an even more extreme result than the Q statistic. The confidence interval for the between-centre variance is fairly wide, especially when considering that $\hat{\sigma}_B^2$ is based

on a large number of centres. However, looking at the sensitivity plot of θ against $\sigma_B^2/(\hat{\sigma}_{Bl}^2+\sigma_B^2)$, changes in the value of σ_B^2 do not affect the estimate of θ to any great extent and certainly not in the region around $\hat{\sigma}_{Bl}^2$ (Figure 14). As a consequence, the likelihood-based confidence interval for θ is apparently identical to the confidence interval derived from the standard random effects method. The interval for θ is approximately symmetric and that for σ_B^2 is also more symmetric than in the previous example, resulting from the larger number of centres involved.

Figure 14: Sensitivity plot showing how the overall difference in mean diastolic blood pressure reduction varies with the between-centre variance



1.78=MLE of between-centre variance

2.3.3 Example 3: An extreme case

The third example is an extreme, and perhaps rather artificial, meta-analysis where there are only two studies to be combined. The two trials investigate the effect of aspirin in the primary prevention of the incidence of stroke, myocardial infarction and

other vascular diseases. Both trials were carried out in a population of male doctors, with one taking place in the U.K. [71] and the other, which was a component of the Physicians Health Survey, in the U.S.A. [80]. Both were randomised controlled trials which compared aspirin (500mg/day in Britain and 325mg/day in U.S.A.) with a placebo. The trial in the U.S.A. had 22,071 participants in the cardiovascular component, whereas the British study was smaller with 5,139 participants. The endpoint considered here is the incidence of non-fatal myocardial infarction (MI) (Table 11) as this provides a situation where there is considerable heterogeneity between the treatment effects (i.e. the log odds ratios) in the two trials. The results used here are those published in the more recent overview of randomised trials of antiplatelet therapy [64], rather than those in the original papers themselves. These two trials are the only two low risk (primary prevention) trials for which the outcome of non-fatal myocardial infarction was reported.

Table 11: Results for two trials of the effect of aspirin in the primary prevention of non-fatal myocardial infarction (MI)

Trial	Number of non-fatal MIs/Total number of patients		Odds ratio	95% C.I.
	Treated	Control	$(e^{\hat{\theta}_i})$	for e^{θ_i}
UK	87/3429(2.5%)	38/1710(2.2%)	1.15	(0.78,1.68)
US	129/11037(1.2%)	211/11034(1.9%)	0.61	(0.49,0.76)

In a fixed effect analysis on these data, there is a significant treatment benefit from taking aspirin (Table 12). However, when using either of the random effects models, the effect observed is no longer significant at the 5% level as both intervals are much wider and include unity. The confidence interval for the overall treatment effect from the likelihood model is also considerably wider than that from the standard random effects model. These results indicate that no conclusion can be reached from these data alone concerning the benefit of aspirin in terms of the risk of non-fatal

myocardial infarction.

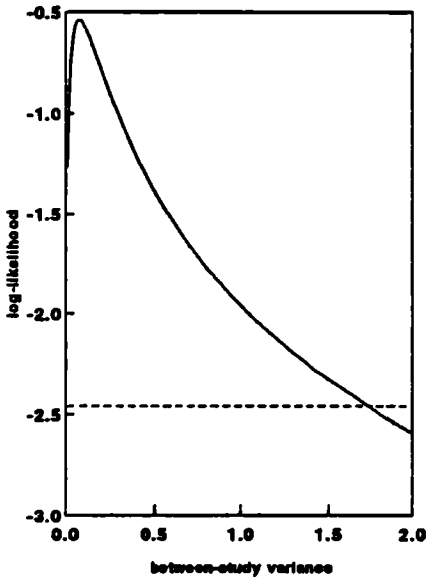
As would be expected the confidence interval for σ_B^2, being based on only two observations, is extremely wide. This is due to the gradual decrease of the profile likelihood for values of σ_B^2 greater than the MLE (Figure 15). This uncertainty in the estimate of σ_B^2, together with the fact that the range of values of σ_B^2 included in the confidence interval lead to a large range of possible estimates of θ (between 0.71 and 0.825 approximately) (Figure 16), results in the large increase in width in the confidence interval for θ. This situation contrasts with that in the previous example where, although the confidence interval for σ_B^2 is quite wide, the specific value that σ_B^2 takes does not influence the overall estimate of the treatment effect. This example also contrasts with the diuretics trials example, since the value of $\hat{\theta}$ is rapidly changing in the region around $\hat{\sigma}_{Bl}^2$ (Figure 16).

The one-sided likelihood ratio test for the null hypothesis of homogeneity $H_0 : \sigma_B^2 = 0$, gives a p-value 0.034, while the test for heterogeneity using Q gives $p \simeq 0.005$ (Q=7.86, χ_1^2) indicating strong evidence of heterogeneity. This, again, perhaps surprisingly, suggests that Q has greater power than the likelihood ratio test.

Table 12: Comparison of the results from three meta-analysis methods for the aspirin trials data

Method	Estimate of between-study variance ($\hat{\sigma}_B^2$)	95% C.I. for σ_B^2	Estimate of overall odds ratio ($e^{\hat{\theta}}$)	95% C.I. for e^{θ}
Fixed effect	0.00	-	0.71	(0.59,0.86)
Random effects				
Standard	0.18	-	0.82	(0.44,1.52)
Likelihood	0.07	(0.00,1.73)	0.80	(0.39,1.78)

Figure 15: Profile likelihood for the between-study variance for the aspirin trials meta-analysis

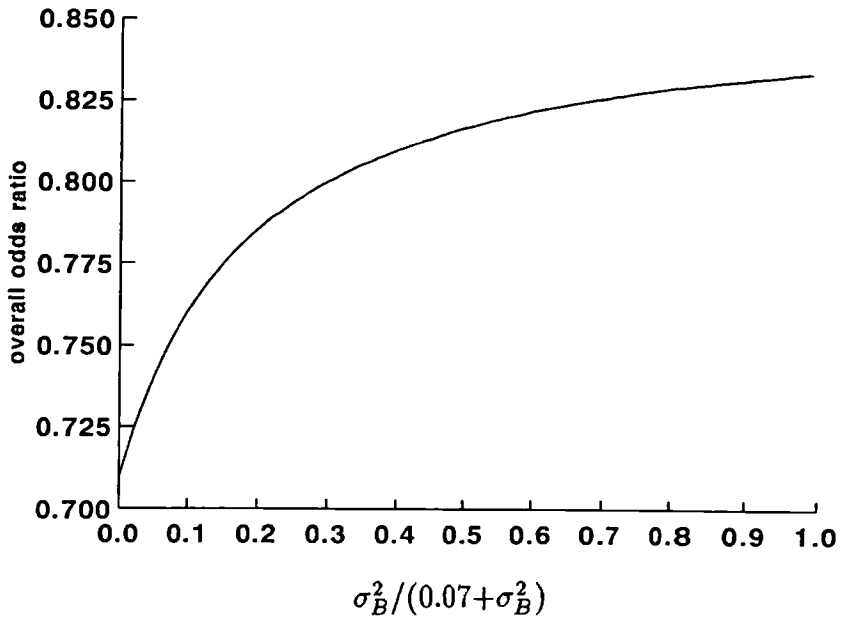


Key

- - maximum log-likelihood − 1.92

0.07=MLE of between-study variance

Figure 16: Sensitivity plot showing how the overall odds ratio varies with the between-study variance for the aspirin trials meta-analysis



0.07=MLE of between-study variance

2.3.4 Use of the information matrix

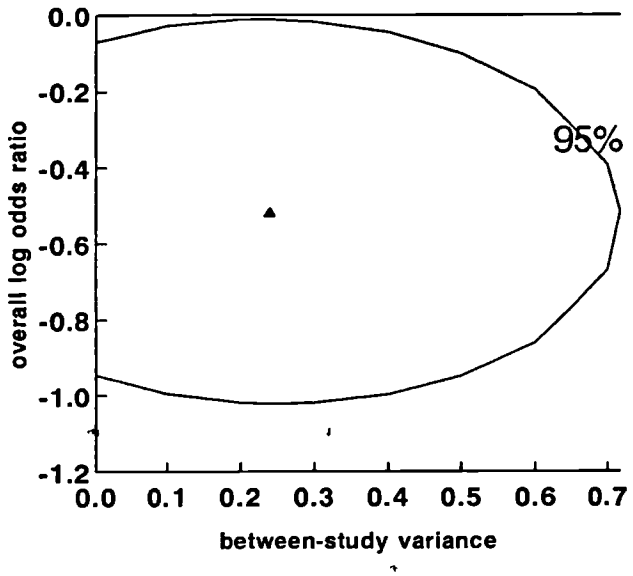
Approximations to the profile likelihoods may be obtained using quadratic curves derived from the asymptotic variance-covariance matrix for $\Phi = (\theta, \sigma_B^2)^T$. Tests, based on this approximation, of the null hypothesis of no treatment effect may also be derived as shown for the case of continuous outcome measures by Rosner [81]. In order to take some account of the fact that σ_B^2 is estimated from a finite number of trials, Rosner [81] suggested that for finite samples the test statistic may be better approximated by a t-distribution than by the N(0,1) distribution. However, as has been shown, the widening of the confidence interval depends more on the strength of the relationship between σ_B^2 and $\hat{\theta}_l$ than it does on simply the number of trials involved and the precision of $\hat{\sigma}_{Bl}^2$. The method using quadratic curves is investigated here so that the resulting intervals can be compared with the likelihood based ones in order to see if the approximation is reasonable in the meta-analysis case.

For a single parameter, the asymptotic distribution of $-2\{$difference in log-likelihood$\}$ tends to that of $(\theta - \hat{\theta}_l)^2 I(\theta)$ [79]. But asymptotically the distribution of θ tends to $N(\theta, I^{-1}(\theta))$, where $I^{-1}(\theta)$ is the information matrix which is given by $E[(-\partial^2 l(\theta)/\partial\theta^2)]$, so $[(\theta - \hat{\theta}_l)/\sqrt{1/I(\theta)}]^2 = (\theta - \hat{\theta}_l)^2 I(\theta)$ tends towards a χ_1^2 distribution. If the expectation cannot be taken algebraically, then $I(\theta)$ may be replaced by the observed information $I^*(\theta)$, which is given by $(-\partial^2 l(\theta)/\partial\theta^2)$ evaluated at $\theta{=}\hat{\theta}_l$. In the case of two parameters, the multivariable Taylor expansion is used to obtain the equivalent result for Φ. It can be shown that the distribution of $-2\{l(\Phi) - l(\hat{\Phi})\}$ tends asymptotically to that of $-(\Phi - \hat{\Phi})^T I(\Phi)(\Phi - \hat{\Phi})$ and that the distribution of $(\Phi - \hat{\Phi})^T I(\Phi)(\Phi - \hat{\Phi})$ tends to a χ_2^2 distribution [79]. This means that the likelihood surface is approximated by a quadratic for both parameters of the model. The values of Φ which satisfy the equation

$$(\Phi - \hat{\Phi})^T I^*(\hat{\Phi})(\Phi - \hat{\Phi}) = 5.991 \qquad (29)$$

where 5.991 is the 95% point on the χ^2_2 distribution, give the 95% confidence region for the two parameters jointly, which may again be displayed on a contour plot. The quadratic approximation would not appear to be very sensible, however, since by approximating the distribution to an ellipse, negative values of σ^2_B can be obtained and the possible asymmetric nature of the likelihood surface is not taken into account. This happens in the case of the diuretics trials data, where the region contains negative values of σ^2_B, which have no meaning, and hence must be set to zero (Figure 17).

Figure 17: 95% contour of the bivariate distribution of the overall log odds ratio and the between-study variance using the quadratic approximation to the likelihood for the diuretics trials meta-analysis



Key

▲ maximum likelihood

A further disadvantage is that equation (29) is fairly complicated to solve. By multiplying the matrices, (29) becomes

$$(\theta-\hat{\theta}_l)^2\sum_{i=1}^{k}\frac{1}{(v_i+\hat{\sigma}^2_{Bl})}+2(\theta-\hat{\theta}_l)(\sigma^2_B-\hat{\sigma}^2_{Bl})\sum_{i=1}^{k}\frac{(\hat{\theta}_i-\hat{\theta}_l)}{(v_i+\hat{\sigma}^2_{Bl})^2}+(\sigma^2_B-\hat{\sigma}^2_{Bl})^2\sum_{i=1}^{k}\frac{2(\hat{\theta}_i-\hat{\theta}_l)-(v_i+\hat{\sigma}^2_{Bl})}{2(v_i+\hat{\sigma}^2_{Bl})^3} \tag{30}$$

which is equal to 5.991. This expression is solved for θ and σ^2_B to obtain the 95% confidence region for θ and σ^2_B. Hence it is actually more straightforward, as well as more meaningful, to produce confidence intervals for each parameter individually.

If the surface of the log-likelihood is quadratic in both parameters, then the interval for a single parameter may be obtained using the respective entry in the observed formation matrix, $I^{*-1}(\hat{\Phi})$ [82]. Hence, for θ for example, $(\theta-\hat{\theta}_l)^2 I^{*-1}_{11}(\hat{\theta}_l)$, where $I^{*-1}_{11}(\hat{\theta}_l)$ is given by $I^{*}_{11}-I^{*}_{12}I^{*-1}_{22}I^{*}_{21}$ and I^{*}_{ij} is the entry (i,j) in the 2x2 variance-covariance matrix, has a χ^2_1 distribution. Alternatively, but equivalently, $(\theta-\hat{\theta}_l)\sqrt{I^{*-1}_{11}(\hat{\theta}_l)}$ has a standard normal distribution and hence values of θ satisfying

$$(\theta-\hat{\theta})\sqrt{I^{*-1}_{11}(\hat{\theta}_l)}=1.96 \tag{31}$$

provide an approximate 95% confidence interval. Similarly, a confidence interval for σ^2_B can be obtained by using the relevant entry in the inverted information matrix, that is I^{*-1}_{22}. However, in the meta-anlysis case under consideration, the surface of the log-likelihood is certainly not quadratic in both directions and so the result is not strictly valid.

Comparing the two confidence intervals for the overall odds ratio estimate for the diuretics trials data, it can be seen that the interval for e^{θ} derived from the profile likelihood is substantially wider than the approximate interval (Table 13, Figure 18). The 95% confidence interval for σ^2_B using the profile likelihood is much wider than that obtained using the variance from the information matrix (Table 13). That the approximation to the confidence interval for e^{θ} using the quadratic is no different to the interval obtained when using the conventional random effects variance,

Table 13: Confidence intervals for the two estimates comparing the profile likelihood method and the quadratic approximation (diuretics trials data)

Method	95% C.I. for the between-study variance	95% C.I. for the overall odds ratio
Profile likelihood	(0.027,1.130)	(0.374,0.953)
Information matrix	(0.000,0.623)	(0.398,0.893)

$1/\sum_{i=1}^{k}(v_i+\sigma_B^2)^{-1}$, may be explained by the fact that the covariance term in the matrix, $Cov(\hat{\theta}_l, \hat{\sigma}^2_{Bl})$, is very small, and therefore has little impact on the value of $I_{11}^{*-1}(\hat{\theta})$. This implies that the variance for $\hat{\theta}$ is equal to $1/\sum_{i=1}^{k}(v_i+\sigma_B^2)^{-1}$ if it is assumed that the two off-diagonal elements of $I^*(\hat{\Phi})$, that is the covariance terms, are zero. Similarly, an approximate variance for σ_B^2 can then be obtained and is given by $1/\sum_{i=1}^{k}(2(\hat{\theta}_i-\hat{\theta}_l)^2-(v_i+\hat{\sigma}^2_{Bl}))/(2(v_i+\hat{\sigma}^2_{Bl}))$.

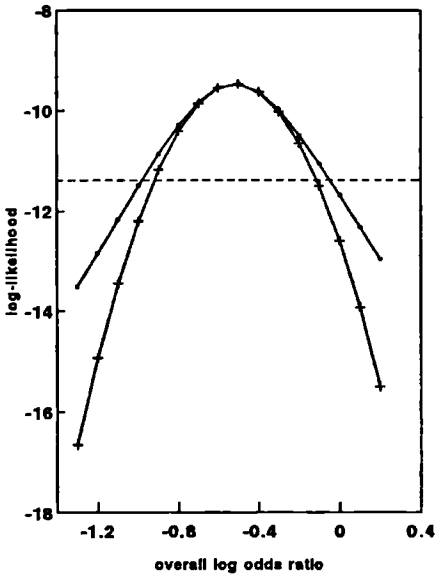
It is possible to use a transformation of σ_B^2 in order to obtain a profile log-likelihood which is a better approximation to a quadratic. In this example, the transformation $\ln(\sigma_B^2)$ does improve the quadratic nature of the profile log-likelihood, but the problem with the log transformation is that it forces all values of σ_B^2 to be greater than zero. The likelihood is certainly more quadratic in shape (Figure 19), but the transformation skews it slightly in the other direction.

The information matrix must be recalculated for the parameters θ and $\ln(\sigma_B^2)$ to obtain the results and hence the reparameterisation gives a log-likelihood of

$$l(\Phi) = -\frac{1}{2}\sum_{i=1}^{n} ln2\pi(v_i + e^{ln\sigma_B^2}) - \frac{1}{2}\sum_{i=1}^{n}\frac{(\hat{\theta}_i-\theta)^2}{(v_i+e^{ln\sigma_B^2})} \qquad (32)$$

Then by making the simplifying substitution, $e^{ln\sigma_B^2} = e^a$, $l(\Phi)$ in (32) can be rewritten as

Figure 18: Profile likelihood for the overall log odds ratio compared to the quadratic approximation to the likelihood for the diuretics trials meta-analysis



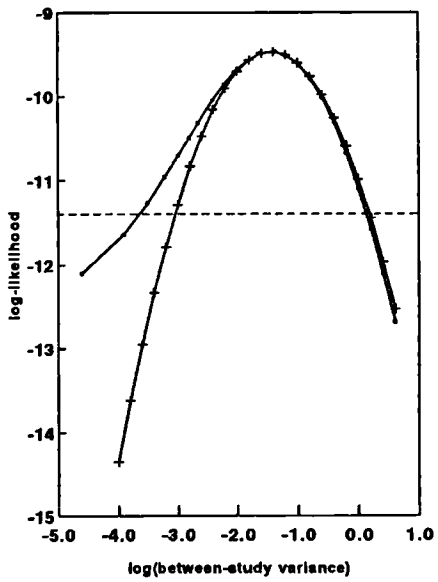
Key

- - maximum log-likelihood − 1.92

-•- profile likelihood

-+- quadratic curve

Figure 19: Profile likelihood for the log of the between-study variance compared to the quadratic approximation to the likelihood for the diuretics trials meta-analysis



Key

- - maximum log-likelihood − 1.92

-•-profile likelihood

-+- quadratic curve

$$l(\Phi) = \frac{1}{2}\sum_{i=1}^{n} ln2\pi(v_i + e^a) - \frac{1}{2}\sum_{i=1}^{n} \frac{(\hat{\theta}_i - \theta)^2}{(v_i + e^a)} \tag{33}$$

Since $I^*(\Phi) = -\partial^2 l/\partial\Phi^2$,

$$I^*(\Phi) = \begin{pmatrix} \sum_{i=1}^{n} \frac{1}{(v_i+\sigma_B^2)} & \sum_{i=1}^{n} \frac{(\hat{\theta}_i-\theta)\sigma_B^2}{(v_i+\sigma_B^2)^2} \\ \sum_{i=1}^{n} \frac{(\hat{\theta}_i-\theta)\sigma_B^2}{(v_i+\sigma_B^2)^2} & \frac{1}{2}\sum_{i=1}^{n} \frac{v_i\sigma_B^2(v_i+\sigma_B^2)^2-(\hat{\theta}_i-\theta)^2(v_i^2+\sigma_B^4)\sigma_B^2}{(v_i+\sigma_B^2)^4} \end{pmatrix} \tag{34}$$

and it follows that the expression for $(\Phi - \hat{\Phi})I^*(\hat{\Phi})(\Phi - \hat{\Phi}) = 5.991$, similar to (30) can be obtained. Furthermore, confidence intervals for θ and $\ln(\sigma_B^2)$ can be calculated (Table 14) in the same way as illustrated in (31).

Table 14: Information matrix-based confidence intervals using θ and $\ln(\sigma_B^2)$

Parameter	95% C.I. of parameter
Overall log odds ratio (θ)	(0.389,0.893)
Log between-study variance ($ln(\sigma_B^2)$)	(-3.042,0.176)
Between-study variance (σ_B^2)	(0.048,1.193)

Comparing these results with those obtained without the transformation the confidence interval for θ remains unchanged, but the interval for σ_B^2 becomes wider and does not include zero (which it cannot, because of the log transformation). This approximation now happens to agree reasonably well with the interval obtained directly from the profile likelihood. This will not generally be true, but it may be that in certain cases a transformation of the variable will improve the quadratic nature of the log-likelihood. The transformation of one parameter will have no effect on the interval of the other parameter. Hence, the interval for θ remains the same after the transformation of σ_B^2.

For this particular set of data, and indeed in general, it is more sensible to use the likelihood based intervals taken directly from the profile likelihood curve. The quadratic approximations based on the use of the information matrix lead to confidence intervals which are too narrow. Using the approximation was computationally not much quicker than using the curve directly, especially since a data-dependent transformation of σ_B^2 was required to obtain a reasonably sensible approximation. The approximations would be rather better for the multicentre trial (Section 2.3.2), since the joint log-likelihood of θ and σ_B^2 is much more quadratic in shape because of the large number of centres involved. However, it would be completely unrealistic for the case of the aspirin trials meta-analysis (Section 2.3.3), as the likelihood for σ_B^2 is far from quadratic.

2.3.5 Discussion

The fact that the likelihood based confidence intervals are based on the 95% confidence level from the χ_1^2 distribution, which requires the quadratic approximation to hold, means that they too are only approximate confidence intervals. They should therefore not strictly be viewed as 95% confidence intervals in the usual sense, but rather as likelihood support intervals. However, these likelihood support intervals may be interpreted for practical purposes as being approximate confidence intervals, as suggested by Clayton and Hills [83].

The likelihood method presented yields a confidence interval for σ_B^2, so that the precision of $\hat{\sigma}_{Bl}^2$ can be directly judged. Obviously the fewer trials involved, the less precise will be the estimate of σ_B^2. However, even in the second example with 189 'trials', the width of the confidence interval was large. Hence, in any meta-analysis in practice there will be considerable imprecision in estimating σ_B^2. Whether the value of σ_B^2 used substantially affects the overall estimated treatment effect is a separate issue, and can easily be investigated using the sensitivity plot of $\hat{\theta}$ (or $\hat{\theta}_l$) against σ_B^2

(Section 2.1). In practical terms, this issue is of importance when σ_B^2 is imprecisely estimated and when the value of σ_B^2 affects the overall estimate of the treatment effect in the region around the MLE.

It has been noted how in the first and third examples (Sections 2.3.1 and 2.3.3), the likelihood ratio test appears to have less power than the Q test to detect heterogeneity. However in the second example, it has greater power. To understand why this occurs, it is necessary to consider the hypotheses which are being tested. In the case of the test for heterogeneity using Q the null hypothesis is that all individual study estimates are equal $H_0 : \theta_i = \theta$ for all i. This is equivalent to the null hypothesis in the likelihood ratio test which is that the between-study variance is zero $H_0 : \sigma_B^2 = 0$. However, the alternative hypotheses differ between the two tests. The alternative for the Q test is $H_1 : \theta_i \neq \theta$ for at least one i, while that for the likelihood ratio test is more specific, being not only that the between-study variance is greater than zero, but also that the normal random effects model holds under this alternative. Since it is the alternative hypothesis that determines the power of a test, it would not necessarily be expected for these two tests to have the same power. Hence, differences in the p-values obtained may not be that surprising.

It would be expected, that if a set of data does follow a normal random effects model reasonably well, then the likelihood ratio test would be more powerful than the general Q test. However, if the data was not of this form then the likelihood ratio test would lose power due to the alternative hypothesis being inappropriate and the Q test would perhaps be more powerful in such a situation. With only a few points normality is difficult to check and is impossible with only two trials. Normal plots (see Chapter 3) for the first example indicate that the normally distributed random effects model may not be very suitable in this case. Hence, the alternative hypothesis of the likelihood ratio test is inappropriate and so the test will lack power when compared to Q.

Although the likelihood method proposed allows for the estimation of the between-study variance, it still assumes that the individual study variances v_i are known, when in practice they too must be estimated. For continuous outcome measures, a possible alternative approach would be to define each v_i as σ^2/n_i, where σ is a common within-study standard deviation. This is a valid substitution if v_i is approximately proportional to $1/n_i$, that is if the variance of $\hat{\theta}_i$ is only dependent on the number of observations on which it is based. The likelihood, based on the normally distributed random effects model, can then be obtained with σ^2/n_i replacing v_i. Hence, the log-likelihood for θ, σ_B^2 and σ^2 is

$$l_i(\theta, \sigma_B^2, \sigma^2) = -\sum_{i=1}^{k} \frac{1}{2} ln 2\pi((\sigma^2/n_i) + \sigma_B^2) - \sum_{i=1}^{k} \frac{(\theta_i - \theta)^2}{2((\sigma^2/n_i) + \sigma_B^2)} \qquad (35)$$

which contains three unknown parameters, θ, σ_B^2 and σ^2, and since n_i are known, this eliminates the problem of having to assume that the v_i are known. The MLEs for θ, σ_B^2 and σ^2 may then be obtained by standard methodology.

The likelihood method using the marginal distributions of $\hat{\theta}_i$ makes the assumption that the data $\hat{\theta}_i$ are normally distributed as well as the random effects. Hence, for binary outcome measures it does not utilise the exact distribution of each 2x2 table. The approximation to the normal distribution may, therefore, be inadequate in some cases, particularly when the sample sizes are small. The full likelihood, for binary data, includes the exact conditional distribution of each 2x2 frequency table given its margins [45]. If a full likelihood method were pursued, the confidence intervals for the overall treatment effect would be expected to be even wider. This issue is explored in the next section.

The examples presented show that caution is required when interpreting results from the standard meta-analysis methods. Even the confidence interval for the overall treatment effect from the usual random effects model may be too narrow. Certainly,

the confidence intervals from a fixed effect model will tend to be far too narrow when heterogeneity exists. This issue is of practical importance since the increased width of the confidence intervals can severely limit the conclusions that can be drawn from a meta-analysis.

2.4 A Full Likelihood Approach For Binary Outcomes

In a paper considering a bivariate approach to meta-analysis, van Houwelingen, Zwindermann and Stijnen used a likelihood based "Mantel-Haenszel-type" procedure for a fixed effect model and extended this to a random effects model [45]. Section 2.4.1 describes the fixed effect model and Section 2.4.2 the random effects model. Section 2.4.3 then compares the results obtained using the full likelihood method with those obtained using the inverse-variance fixed effect method (Section 1.5.1) and the random effects likelihood method based on the marginal distributions of $\hat{\theta}_i$ (Section 2.2) for two practical examples.

2.4.1 The fixed effect model

The likelihood used in [45], under the homogeneity assumption, was obtained by considering the conditional distribution of the number of events in the control group c_i, given the total number of events in that trial m_{i1} (notation as in Table 2). However, in order that the odds ratios obtained are in the same direction as those in the rest of the thesis, the conditional distribution for the number of events in the treatment group will actually be considered here. In each trial, the conditional distribution of the number of events in the treatment group a_i given m_{i1} is a non-central hypergeometric distribution of the 2x2 table given its marginals [45], and so the likelihood of θ, the log odds ratio, is

$$L_i(\theta) = \frac{\binom{n_{i1}}{a_i}\binom{n_{i2}}{c_i} e^{\theta a_i}}{\sum_{y=y_l}^{y_u} \binom{n_{i1}}{y}\binom{n_{i2}}{a_i + ci - y} e^{\theta y}} \tag{36}$$

where $y_l = max(0, m_{i1} - n_{i2}) \leq y \leq min(n_{i1}, m_{i1}) = y_u$. The denominator of (36) represents the sum over all possible tables where the total number of events is equal to that observed.

The total log-likelihood is the sum of the individual trial likelihoods, so if $l_i(\theta) = lnL_i(\theta)$, then

$$l(\theta) = \sum_{i=1}^{k} l_i(\theta) \tag{37}$$

This total likelihood can therefore be plotted and the MLE of θ obtained directly from the curve, together with $l(\hat{\theta})$ and the 95% confidence interval for θ. The 95% confidence interval is obtained, as in Section 2.2, by taking as the confidence limits the points at the intersection of the likelihood curve and the horizontal line drawn at 3.84/2 (3.84 is the 95% point on the χ_1^2 distribution) units below the maximum.

2.4.2 The random effects model

This likelihood model was extended by van Houwelingen et al. [45] to incorporate random effects. This was done by assuming θ to be random with some distribution G and so θ_i, $i = 1, ..., k$, is a random sample from G, the values of θ_i being unobservable. Then assuming that θ is independent of the sample sizes, the likelihood for trial i is given by

$$L_i(G(\theta)) = \int_\theta L_i(\theta) dG(\theta) \tag{38}$$

and hence the parameters of the distribution G can be estimated by maximising the total log-likelihood of $G(\theta)$,

$$l(G(\theta)) = \sum_{i=1}^{k} l_i(G(\theta)) \tag{39}$$

Either the nonparametric approach of Laird [84] or a parametric approach may be used to estimate the distribution of G. The parametric approach leads to a smoother estimate of G being obtained [45], and van Houwelingen et al. assume a normal distribution as in the marginal likelihood model of Section 2.2. Hence $G \sim N(\mu, \sigma_B^2)$ where μ is now the overall treatment effect and σ_B^2 is the between-study variance and so,

$$L_i(\mu, \sigma_B^2) = \int_\theta L_i(\theta) \frac{1}{\sqrt{2\pi\sigma_B^2}} exp \left\{ \frac{-(\theta - \mu)^2}{2\sigma_B^2} \right\} d\theta \tag{40}$$

where $L_i(\theta)$ is given in (36). The MLEs of μ and σ_B^2 can then be obtained by implementation of the EM algorithm [85]. The EM algorithm consists of two steps, the estimation step (E-step) and the maximisation step (M-step), which are repeated alternately until convergence is achieved. The E-step involves the computation of the sufficient statistics of the θ_i, which are $\sum_{i=1}^{k} \theta_i$ and $\sum_{i=1}^{k} \theta_i^2$ if $\theta_1, ..., \theta_k$ could be observed. Hence, if $f(\theta \mid \mu, \sigma_B^2)$ is the normal density function, then

$$\tilde{\theta}_i = E(\theta_i \mid data, parameters) = \frac{\int \theta L_i(\theta) f(\theta \mid \mu, \sigma_B^2) d\theta}{\int L_i(\theta) f(\theta \mid \mu, \sigma_B^2) d\theta} \tag{41}$$

and

$$\tilde{\tau}_i = E(\theta_i^2 \mid data, parameters) = \frac{\int \theta^2 L_i(\theta) f(\theta \mid \mu, \sigma_B^2) d\theta}{\int L_i(\theta) f(\theta \mid \mu, \sigma_B^2) d\theta} \tag{42}$$

Numerical methods can be used in order to evaluate the integrals given in (41) and (42).

The M-step then consists of calculating the mean and the variance of the distribution as follows.

$$\hat{\mu} = \frac{1}{k}\sum_{i=1}^{k}\tilde{\theta}_i \tag{43}$$

$$\hat{\sigma}^2_{Bv} = \frac{1}{k}\sum_{i=1}^{k}\tilde{\tau}_i - \hat{\mu}^2 \tag{44}$$

A 95% confidence interval may be produced using the profile likelihood (defined in Section 2.2.3) of $\hat{\mu}$. The EM algorithm can again be used in order to obtain the maximum value of σ^2_B for each value of $\hat{\mu}$. Hence, the procedure is exactly the same as described above, except that the value of $\hat{\mu}$ is known in $L_i(G)$. Once again the horizontal line at 1.92 units below the maximum is used and so the 95% confidence interval is defined as in equation (28).

Furthermore, a likelihood ratio test for heterogeneity may be formulated by calculating the test statistic $-2\{l(\hat{\theta}) - l(\hat{\mu}, \hat{\sigma}^2_{Bv})\}$, where $l(\hat{\theta})$ is the maximum likelihood for the homogeneous model, that is where σ^2_B=0, and $l(\hat{\mu}, \hat{\sigma}^2_{Bv})$ is the maximum likelihood for the random effects model. This is equivalent to the test described in Section 2.2.3, except that the full likelihood as opposed to the marginal likelihood is used.

2.4.3 Comparison of results

A Gauss program, provided by van Houwelingen et al. [45], which carries out the parametric likelihood analysis described above, was used in order to obtain results for

both the diuretics trials data (Section 1.3.1) and the aspirin trials data (Section 2.3.3). The results from the random effects methods were then compared with those obtained from the marginal likelihood method (Section 2.2) and the results obtained under the homogeneity assumption were compared with the inverse-variance fixed effect method (Section 1.5.1).

The comparison of the results show that in the case of the diuretics trials meta-analysis, the two fixed effect methods agree closely (Table 15). The differences between the random effects Mantel-Haenszel-type likelihood results and the marginal likelihood results are also very small (Table 16). The 95% confidence interval is slightly wider and the estimate of σ_B^2 is slightly larger in the full likelihood method. An increase in uncertainty is expected since the full likelihood does not make the assumption that the variances of the individual studies are known. However, it is in the smallest studies that the variances are most imprecisely estimated. Hence, such studies take the least weight in a meta-analysis and also have their relative weight determined more by the value of σ_B^2 than by v_i. This means that the additional uncertainty would not be expected to have a great impact on the results, and so pursuing the full likelihood approach may be unnecessarily sophisticated for most purposes.

Table 15: Comparison of results from the fixed effect likelihood based Mantel-Haenszel-type procedure with those from the inverse-variance fixed effect method for the diuretics trials data

Method	Estimate of overall odds ratio	95 % C.I. for overall odds ratio
M-H likelihood	0.66	(0.56,0.79)
Inverse-variance	0.67	(0.56,0.80)

The results for the meta-analysis of the two aspirin trials are also very com-

Table 16: Comparison of results from the random effects likelihood based Mantel-Haenszel-type procedure with those from the marginal likelihood random effects method for the diuretics trials data

Method	Estimate of overall odds ratio	95 % C.I. for overall odds ratio	Estimate of between-study variance
M-H likelihood	0.60	(0.37,0.97)	0.26
Marginal likelihood	0.60	(0.37,0.95)	0.24

parable (Tables 17 and 18). However, surprisingly, the confidence interval for the overall odds ratio is narrower in the case where the full likelihood is used. This may be due to the fact that only two trials are being analysed leading to a peculiarly shaped likelihood curve, particularly away from the maximum. Furthermore, the two likelihood based confidence intervals are both much wider than that obtained using the standard random effects model (Table 12). It is also noticeable that the two likelihood estimates of σ_B^2 agree with each other (0.07 and 0.08), whereas the D&L moment estimate is much larger at 0.18 (Table 12)

Table 17: Comparison of results from the fixed effect likelihood based Mantel-Haenszel-type procedure with those from the inverse-variance fixed effect method for the aspirin trials data

Method	Estimate of overall odds ratio	95 % C.I. for overall odds ratio
M-H likelihood	0.71	(0.60,0.86)
Inverse-variance	0.71	(0.59,0.86)

The likelihood ratio tests based on the full likelihood give a p-value of 0.0025 for the diuretics trials data and one of 0.038 for the aspirin trials data. Both are less powerful than the corresponding test using Q, as were the tests derived from the

Table 18: Comparison of results from the random effects likelihood based Mantel-Haenszel-type procedure with those from the marginal likelihood random effects method for the aspirin trials data

Method	Estimate of overall odds ratio	95 % C.I. for overall odds ratio	Estimate of between-study varaince
M-H likelihood	0.80	(0.40,1.71)	0.08
Marginal likelihood	0.80	(0.39,1.78)	0.07

marginal likelihood model. However, the p-value for the diuretics trials data using the full likelihood is substantially smaller than that using the marginal likelihood. This may be due to the fact that the alternative hypothesis takes a different form, which may, in this example, better represent the actual data.

Hence, in the examples considered so far, there is no clear advantage to be gained from using a Mantel-Haenszel-type procedure as opposed to the marginal likelihood model. Certainly the likelihood method based on the marginal distributions of $\hat{\theta}_i$ provides a good approximation to the full likelihood in circumstances where the number of events in each trial is fairly large. Furthermore, the marginal likelihood approach may also be used to analyse continuous outcome measures as well as binary. However, the great advantage of the full likelihood is when there are studies which have small and, most particularly, zero event rates. This issue is pursued in the next section.

2.5 Dealing with Small Event Rates in Meta-Analyses

The problem of dealing with small event rates and zero event rates in trials included in meta-analyses using binary outcome measures is introduced in Section 2.5.1. Exact methods which are not prone to such problems and are based on the exact conditional

likelihood are described in Section 2.5.2, while a comparison of the methods, using two practical examples, is carried out in Section 2.5.3.

2.5.1 Introduction

When dealing with binary outcomes, if there is a trial included in the meta-analysis which has a zero cell in the 2x2 table, then certain of the standard methods fail, since it becomes impossible to calculate an individual odds ratio $\hat{\theta}_i$ and variance v_i. Individual trial odds ratios and variances must be calculated explicitly in the inverse-variance fixed effect method (Section 1.5.1), the standard random effects method (Section 1.7.1) and the marginal likelihood method (Section 2.2), and hence it is these methods that break down. Even if there are no zero event rates, but the event rates are small, then the asymptotic conditions underlying the standard meta-analysis methods will not hold [60].

One way around the problem is to add 0.5 to each cell of each 2x2 table in the meta-analysis and thus obtain empirical logits [86]. The use of empirical logits ensures that both odds ratio and variance estimates may be obtained in all trials, and will also reduce the bias for small sample sizes [50, 86]. Alternatively, estimation can be based on the exact conditional distribution of the number of events in the treatment group (Section 2.4). It was shown in Section 2.4 how van Houwelingen et al. [45] used the exact distribution of $\hat{\theta}_i$ and likelihood methodology to obtain estimates of the overall treatment effect under both a fixed effect and a random effects model.

2.5.2 Conditional likelihood model

Similarly, although only for homogeneous data, the computer package StatXact [87] uses the conditional likelihood of the sufficient statistic $S = A_1 + A_2 + ... + A_k$, where A_i is the true number of events in the treatment group of study i, to produce exact

estimates and confidence limits [87]. The total log-likelihood for all k tables is given by the sum of the k individual log-likelihoods and so the conditional distribution of the total number of events S is considered. Hence, if the observed value of S is $s = \sum_{i=1}^{k} a_i$, then the conditional distribution of S, that is the probability of observing a total of S events given all possible combinations of the k tables with the observed marginals, can be written as follows [88]

$$P(S = s \mid \psi) = \frac{c_s \psi^s}{\sum_{y=y_L}^{y_U} c_y \psi^y} \tag{45}$$

where $\psi = e^{\theta}$

$$c_s = \sum_{\tau \in \Omega(s)} \prod_{i=1}^{k} \binom{n_{i1}}{a_i} \binom{n_{i2}}{c_i}$$

$$\Omega(s) = \{\tau \in \Omega : y_1 + y_2 + \ldots + y_k = s\}$$

$$y_L = \sum_{i=1}^{k} max(0, m_{i1} - n_{i2})$$

$$y_U = \sum_{i=1}^{k} min(n_{i1}, m_{i1})$$

A test of $\psi = e^{\theta}$=1 is based on this conditional distribution and an exact confidence interval may be constructed by inverting this test [59]. Specifically, an exact 100 (1 − α) % confidence interval for θ is given by $\{\psi_*(s), \psi^*(s)\}$, where $\psi_*(s)$ is such that

$$\psi_*(s) = 0 \quad if \quad s = y_L \tag{46}$$

$$P(S \geq s \mid \psi_*(s)) = \alpha/2 \quad if \quad y_L < s < y_U \tag{47}$$

$$P(S = s \mid \psi_*(s)) = \alpha \quad if \quad s = y_U \tag{48}$$

and $\psi^*(s)$ is such that

$$P(S = s \mid \psi^*(s)) = \alpha \quad if \quad s = y_L \tag{49}$$

$$P(S \leq s \mid \psi^*(s)) = \alpha/2 \quad if \quad y_L < s < y_U \tag{50}$$

$$\psi^*(s) = \infty \quad if \quad s = y_U \tag{51}$$

Hence, if the observed value of s is equal to the lower limit y_L, then the data support a lower confidence bound of zero, and so all of the error rate may be used computing the upper bound. Similarly if y_U is observed the entire error rate can be used in calculating the lower bound as the upper bound is ∞. Mehta, Patel and Gray [89] developed a numerical algorithm for computing this exact confidence interval and this is implemented by StatXact. The main problem is the time taken to compute all the different possible values of c_s and once this is done the confidence interval is found easily using a binary search of all the probabilities calculated. However, due to the discreteness of the distribution of S, the above exact confidence interval is conservative, with the probability usually being less than α [87]. StatXact, therefore, also produces mid-p adjusted intervals [87] which reduce the conservativeness. To calculate these corrected intervals if $y_L < s < y_U$ equation (47) is replaced by

$$\frac{1}{2}P(S = s \mid \psi_*(s)) + P(S > s \mid \psi_*(s)) = \alpha/2 \tag{52}$$

and equation (50) is replaced by

$$\frac{1}{2}P(S = s \mid \psi^*(s)) + P(S < s \mid \psi^*(s)) = \alpha/2 \tag{53}$$

2.5.3 Examples

The problem of analysing a meta-analysis when there are small numbers of events was investigated in practical terms by means of two examples. In the diuretics trials meta-analysis, the number of stillbirths was recorded in eight of the nine trials and the total number of such events was small and was actually zero in some groups (Table 19). There were no stillbirths in either group in trials 8 and 9 and as these

trials contribute no information to the analysis the meta-analysis results are actually based on six trials only.

Table 19: Number of stillbirths recorded for the six trials of diuretics taken during pregnancy which contribute information to this outcome

Trial	Stillbirths/Total number of patients		Odds Ratio
	Treated	Control	
1	1/131(0.8%)	2/136(1.4%)	0.52
2	3/335(0.9%)	2/110(1.8%)	0.49
3	1/57(1.8%)	1/48(2.1%)	0.84
4	0/34(0.0%)	1/40(2.5%)	0.00
5	6/1011(0.6%)	5/760(0.7%)	0.90
7	6/1370(1.2%)	9/1336(0.7%)	0.65

These data were analysed using the likelihood based Mantel-Haenszel-type procedures, StatXact and also using the Peto and Mantel-Haenszel fixed effect methods. Results may be calculated by all these methods when there are zero event rates, although the asymptotics of the Peto method may not be very good with small sample sizes. The Mantel-Haenszel estimator is known to be robust in cases where there are small samples [50]. The results obtained from these methods were also compared with those from the inverse-variance fixed effect method and the standard random effects method based on empirical logits.

In this example $\hat{\sigma}^2_B$=0 and $\hat{\sigma}^2_{Bv}$=0 and hence the fixed effect and random effects methods produce the same results and so only the fixed effect results are presented here (Table 20). The various confidence intervals do differ slightly (Table 20) and it can be seen that the exact confidence interval from StatXact is wider than all the others. This illustrates the conservative nature of the exact confidence interval, while the corrected mid-p interval is in line with the other fixed effect intervals. The

confidence interval for the inverse-variance method is slightly narrower than all the others. However, in this example there is no evidence that the asymptotic methods produce inappropriate or unreliable results.

Table 20: Results for the outcome of stillbirths in the diuretics trials meta-analysis using several different fixed effect methods

Method	Estimate of overall odds ratio	95% C.I. for overall odds ratio
Likelihood	0.68	(0.35,1.32)
StatXact (exact C.I.)	0.68	(0.33,1.38)
StatXact (mid-p C.I. *)		(0.35,1.32)
Mantel-Haenszel	0.68	(0.35,1.31)
Peto	0.68	(0.35,1.31)
Inverse-variance (+0.5)	0.69	(0.38,1.28)

* the mid-p method is only an adjustment to the confidence interval

An example with small numbers of events and statistically significant heterogeneity was then considered. The data are taken from a published meta-analysis of the efficacy of BCG vaccine in the prevention of tuberculosis (TB) [90]. The results of the 7 clinical trials which provided information on the number of TB deaths in the vaccinated and unvaccinated groups were used, as the number of deaths was small and there were two groups where no TB deaths occurred.

The results from each trial vary considerably, as can be seen from Table 21, and this is probably due to the fact that the populations and geographical locations in which the trials were carried out vary enormously. Hence, it is not surprising to find that heterogeneity exists and $\hat{\sigma}_B^2$=0.26 by the D&L moment estimator based on the use of empirical logits, and $\hat{\sigma}_{Bv}^2$=0.33 by the full likelihood method. All results, using both fixed effect and random effects methods, show a significant reduction in

Table 21: Results for 7 clinical trials looking at the efficacy of the BCG vaccine in relation to TB deaths

Trial	TB deaths/Total number of patients		Odds Ratio
	Vaccinated	Unvaccinated	
Aronson, 1948	0/123(0.00%)	4/139(2.88%)	0.00
Ferguson, 1949	2/136(1.47%)	9/303(2.97%)	0.50
Rosenthal, 1960	0/231(0.00%)	4/220(1.82%)	0.00
Rosenthal, 1961	1/1716(0.06%)	6/1665(0.36%)	0.16
Comstock, 1974	8/50634(0.02%)	12/27338(0.04%)	0.36
Aronson, 1958	13/1541(0.84%)	68/1451(4.69%)	0.18
Levine, 1948	8/566(1.41%)	8/528(1.52%)	0.93

the number of TB deaths in the vaccinated groups (Table 22). The two random effects methods of course provide confidence intervals which are wider than those for the fixed effect methods. The conservativeness of the exact confidence interval from StatXact is again evident.

It is noticeable that the three exact methods and the standard Mantel-Haenszel estimate produce values for θ which agree very well. However, the asymptotic methods, and particularly the methods based on empirical logits, produce larger estimates. The use here of emirical logits does appear to slightly affect the results, as both the inverse-variance fixed effect method and the standard random effects method produce the two largest estimates of treatment effect and are more similar to each other, even though different models are being assumed, than they are to any of the other estimates. However, use of empirical logits only produce a shift in the confidence intervals, as a consequence of a different estimate, rather than a change in the width. Overall, since there is some disagreement between estimates and heterogeneity is present, the exact random effects likelihood estimate may be preferable in this case.

However, the same conclusion of efficacy of the vaccine against TB would be drawn from all methods.

Table 22: Results for the efficacy of BCG vaccine in the prevention of TB deaths using different meta-analysis methods

Method	Estimate of overall odds ratio	95% C.I. for overall odds ratio	Estimate of between-study variance
Fixed effect methods			
Likelihood	0.24	(0.16,0.35)	
StatXact exact	0.24	(0.16,0.37)	
StatXact mid-p		(0.16,0.36)	
Mantel-Haenszel	0.24	(0.16,0.36)	
Peto	0.27	(0.20,0.38)	
Inverse-variance (+0.5)	0.28	(0.19,0.41)	
Random effects methods			
Likelihood	0.23	(0.07,0.46)	0.33
Standard (+0.5)	0.29	(0.16,0.54)	0.26

Neither of the practical examples considered here has shown the asymptotic methods to be completely unreliable. Hence, in most practical situations, this suggests that any of the above methods may be adequate. However, it may be the case that the asymptotic methods become less reliable when the total numbers in each trial, as well as the event rates, become small. Further examples would be necessary to investigate this issue more fully.

2.6 Bayesian Approach to Meta-Analysis

As an alternative to the classical statistics approach, meta-analysis may be viewed from a Bayesian perspective. Although a more detailed consideration of the latter approach is outside the scope of this thesis, a brief review of Bayesian and empirical Bayes methods is included here. Such Bayesian methods use ideas related to those used in the likelihood based random effects approaches described in previous sections of this chapter. Empirical Bayes methods are considered first, the concept being introduced in Section 2.6.1 and the methodology described in Section 2.6.2. An example, using the diuretics trials data, is presented in Section 2.6.3. The literature relating to a fully Bayesian approach to meta-analysis is then reviewed in Section 2.6.4.

2.6.1 Introduction to empirical Bayes

If heterogeneity is present in a meta-analysis, then presenting only the overall estimate of treatment effect and its variance may not be sensible from a clinical point of view, since it does not provide an idea of the actual range of estimates that the trials produce. However, it is difficult to compare the initial observations $\hat{\theta}_i$ as the precision related to each estimate can vary quite considerably, as it does in the diuretics trials data for example. The calculation of empirical Bayes estimates for each trial means that the individual trial estimates become more directly comparable [31], as they will be of a more similar precision.

The general idea behind empirical Bayes estimation is that of shrinkage, whereby each individual observation is pulled in towards the overall mean value. This concept is intuitively appealing because it means that the outlying estimates, and particularly those with large variances, which appear unlikely in the context of the full set of data, can be brought into line with the other evidence. Hence, in a meta-analysis, a new estimate of treatment effect is obtained for each trial which also takes into account the

combined information from all of the other trials. By incorporating these extra data, the empirical Bayes estimates are more precise than the original observed estimates of treatment effect. The ranking of the empirical Bayes estimates will on average more resemble the ranking of the true estimates more than the original observed values [31]. Furthermore, when considered as a whole, the initial observed treatment effects $\hat{\theta}_i$ will be biased, even though they are individually unbiased [31]. For example, the largest observation is likely to be an overestimate and the smallest an underestimate of the true treatment effect, where the greater the sampling variability, the more likely the under- or over-estimation. Hence, empirical Bayes estimates have certain advantages over the simple observations of treatment effect in a meta-analysis.

Hedges and Olkin [39] and Stijnen and van Houwelingen [31] proposed empirical Bayes estimation for random effects models. Hedges and Olkin made the distributional assumption of normality for the random effects and used the EM algorithm [85] in order to obtain estimates of θ, σ_B^2 and θ_i for i=1,...,k simultaneously. Stijnen and van Houwelingen [31] proposed methods where the distribution of the random effects were both parametric and nonparametric. Furthermore, Morris [91] used empirical Bayes inference to investigate the general problem of interpreting multiple estimates of the same quantity, but the methodology is easily adapted to the case of meta-anlaysis.

2.6.2 Empirical Bayes methods

The setting considered here is that of the normally distributed random effects model. In a Bayesian context, θ_i can be thought of as having a prior distribution which is normal with mean θ and variance σ_B^2 [31], while the observed data $\hat{\theta}_i$ is also normal with mean θ_i and variance v_i. The Bayes estimates can therefore be considered as those obtained from the posterior distribution of θ_i. The posterior distribution is obtained using standard Bayes theory, which is straightforward in this case due to

the normality assumption being made, and is found to be normal with mean $\tilde{\theta}_i$, and variance $\tilde{v}_i$, given by [31, 77, 91]:

$$\tilde{\theta}_i = \hat{\theta}_i \frac{\sigma_B^2}{(v_i + \sigma_B^2)} + \theta \frac{v_i}{(v_i + \sigma_B^2)} \tag{54}$$

$$\tilde{v}_i = \frac{\sigma_B^2 v_i}{(v_i + \sigma_B^2)} \tag{55}$$

In the empirical Bayes case θ and σ_B^2 are then estimated from the data, rather than their distributions being obtained through subjective judgement or prior knowledge as they would be in the fully Bayesian case. The form of equation (54) may be simplified by defining the 'shrinkage factor' B_i, where $B_i = v_i/(v_i + \sigma_B^2)$ [91], that is the proportion by which the estimate is shrunk towards the mean, and so

$$\tilde{\theta}_i = (1 - B_i)\hat{\theta}_i + B_i\theta \tag{56}$$

Obviously, estimates of θ and σ_B^2 are required in order to obtain the empirical Bayes estimates. Hence, the empirical Bayes estimate for the treatment effect in each trial $\tilde{\theta}_i$ consists of a linear combination of the individual observed estimate $\hat{\theta}_i$ and the random effects estimate of the overall mean. The resulting empirical Bayes estimate thus depends on the relative size of the within-study and the between-study variance for each particular trial. Stijnen and van Houwelingen [31] suggest the use of the weighted mean to estimate θ and the D&L moment estimator of σ_B^2. The estimates can also be obtained either using maximum likelihood methods, described in Section 2.2.1, or by an alternative method of moments suggested by Maritz and Lwin [77]. These alternative moment estimators are obtained by equating the estimates of θ and $var(\theta)$ with their expectations. Hence, the following expressions are produced:

$$\hat{\theta} = \frac{\sum_{i=1}^{k} \frac{\hat{\theta}_i}{(v_i+\sigma_B^2)}}{\sum_{i=1}^{k} \frac{1}{(v_i+\sigma_B^2)}} \tag{57}$$

$$\sum_{i=1}^{k} \frac{(v_i + \hat{\sigma}_B^2)}{(\hat{\theta}_i - \hat{\theta})^2} = k \tag{58}$$

The moment estimator of the overall treatment effect θ is the same as the MLE and is again simply the weighted average taking σ_B^2 into account. By subtracting $\sum_{i=1}^{k} v_i/(v_i + \hat{\sigma}_B^2)$ from both sides of (58) and rearranging, a form is obtained which is convenient for iteration,

$$\hat{\sigma}_B^2 = \frac{\sum_{i=1}^{k} \frac{(\hat{\theta}_i-\hat{\theta})^2-v_i}{(v_i+\hat{\sigma}_B^2)}}{\sum_{i=1}^{k} \frac{1}{(v_i+\hat{\sigma}_B^2)}} \tag{59}$$

This expression for $\hat{\sigma}_B^2$ is very similar to that used for the maximum likelihood method, given in (23), except that a squared term is missing in the denominator of each sum. Equations (57) and (59) can now be solved iteratively in the same manner as the maximum likelihood equations. Although equations (57) and (59) were derived using the method given by Maritz and Lwin [77], the actual equations given in this text were found to be incorrect and so the necessary corrections were made.

Both the maximum likelihood estimates and the alternative method of moments estimates of θ and σ_B^2 were used in the analysis of the diuretics trials data as an example. This then allowed the individual empirical Bayes estimates to be found together with their variances and shrinkage factors. The two sets of results (using different estimators of σ_B^2) were compared and the empirical Bayes estimates were also compared with the original observed odds ratios.

2.6.3 Results

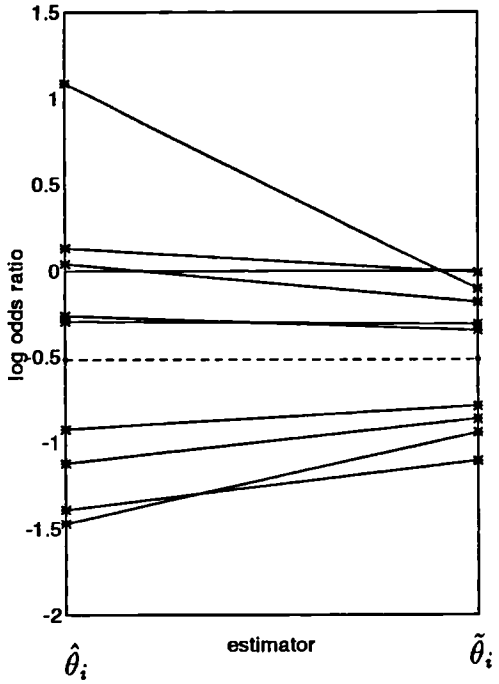
The effect on the individual study estimates of using the empirical Bayes method can clearly be seen, in that the individual log odds ratios are pulled towards the overall weighted mean (Figures 20 and 21, Tables 23 and 24). The empirical Bayes estimates based on the maximum likelihood estimates of θ and σ_B^2 are all below zero (Figure 20), indicating that each trial produces an estimate which may be considered as being consistent with a treatment benefit.

The moment estimate of σ_B^2 (Table 24) is larger than the corresponding MLE (Table 23). This therefore has the effect of making the shrinkage factors smaller which means that the estimates are not pulled towards the overall mean by such a large amount (Figure 21). The two separate estimates of the θ are, however, almost equal.

Both sets of posterior variances are smaller than their corresponding v_i (Tables 23 and 24), which is an implicit characteristic of the empirical Bayes estimates. This reduction is due to the extra data being incorporated into the estimates, thus making them more precise. The variances of the estimates using the moment estimators are larger than those obtained using maximum likelihood estimators. Again this is because the moment estimate of the between-study variance is larger. A problem, however, with the empirical Bayes method is that the estimate of the variance σ_B^2 from the data will rarely be precise (Section 2.2) and hence there is a danger of underestimating the uncertainty in the resulting inferences [40].

The amount by which the individual trial odds ratios move depends both on the particular within-study variance v_i and the absolute distance, $\mid \hat{\theta}_i - \hat{\theta}_r \mid$, of the individual estimate from the mean. The estimate from trial 8, which has a very large variance in comparison to the MLE of the between-study variance ($v_i = 3\hat{\sigma}_{BI}^2$ approximately) is pulled in very considerably. This is because the large v_i causes B_i

Figure 20: Comparison of the observed log odds ratios in each trial with the corresponding empirical Bayes estimates for the diuretics trials meta-analysis using maximum likelihood methods to obtain θ and σ_B^2



Key

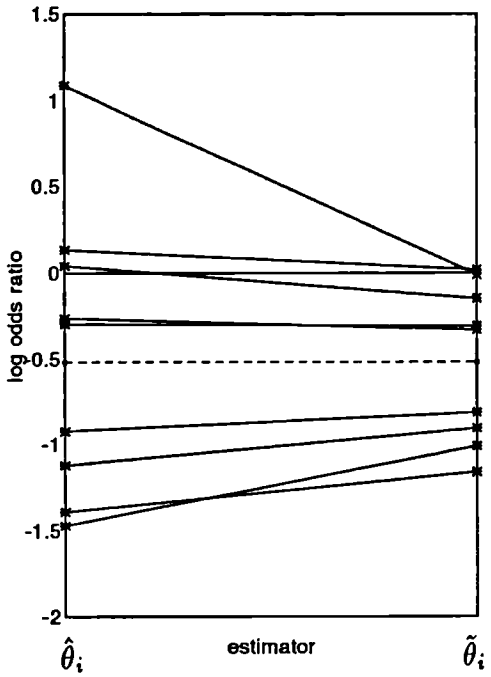
$\hat{\theta}_i$ observed log odds ratio

$\tilde{\theta}_i$ empirical Bayes estimate of log odds ratio

- - Estimate of overall log odds ratio ($\hat{\theta}_l$)=−0.5171

Estimate of between-study variance ($\hat{\sigma}^2_{Bl}$)=0.2386

Figure 21: Comparison of the observed log odds ratios in each trial with the corresponding empirical Bayes estimates for the diuretics trials meta-analysis using moment estimators to obtain θ and σ_B^2



Key

$\hat{\theta}_i$ observed log odds ratio

$\tilde{\theta}_i$ empirical Bayes estimate of log odds ratio

- - Estimate of overall log odds ratio ($\hat{\theta}$)=−0.5181

Estimate of between-study variance ($\hat{\sigma}_B^2$)=0.3170

Table 23: Empirical Bayes estimates using maximum likelihood estimates of the overall treatment effect θ and the between-study variance σ_B^2 (diuretics trials meta-analysis)

Estimate of overall treatment effect $(\hat{\theta}_l)$=−0.5171

Estimate of between-study variance $(\hat{\sigma}_{Bl}^2)$=0.2386

Trial	Observed log odds ratio $(\hat{\theta}_i)$	Variance of $\hat{\theta}_i$	Empirical Bayes odds ratio $(\tilde{\theta}_i)$	Variance of $\tilde{\theta}_i$	Shrinkage factor (B_i)
1	0.04185	0.159601	−0.18218	0.095632	0.4008
2	−0.92367	0.117737	−0.78934	0.078836	0.3304
3	−1.12214	0.178018	−0.86361	0.101952	0.4273
4	−1.47331	0.298927	−0.94155	0.132689	0.5561
5	−1.39102	0.114285	−1.10800	0.077273	0.3239
6	−0.29698	0.014634	−0.30961	0.013788	0.0579
7	−0.26155	0.120687	−0.34739	0.080148	0.3359
8	1.08876	0.686372	−0.10286	0.177052	0.7421
9	0.13531	0.067877	−0.00919	0.052844	0.2215

Table 24: Empirical Bayes estimates using moment estimates of the overall treatment effect θ and the between-study variance σ_B^2 (diuretics trials meta-analysis)

Estimate of overall treatment effect $(\hat{\theta}_r)$=−0.5181

Estimate of between-study variance $(\hat{\sigma}_B^2)$=0.3170

Trial	Observed log odds ratio $(\hat{\theta}_i)$	Variance of $\hat{\theta}_i$	Empirical Bayes odds ratio $(\tilde{\theta}_i)$	Variance of $\tilde{\theta}_i$	Shrinkage factor (B_i)
1	0.04185	0.159601	−0.14566	0.106163	0.3349
2	−0.82367	0.117737	−0.81382	0.085844	0.2708
3	−1.12214	0.178018	−0.90491	0.113993	0.3596
4	−1.47331	0.298927	−1.00972	0.153840	0.4853
5	−1.39102	0.114285	−1.15970	0.084005	0.2650
6	−0.29689	0.014634	−0.30665	0.013980	0.0441
7	−0.26155	0.120687	−0.33229	0.087409	0.2757
8	1.08876	0.686372	−0.01043	0.216860	0.6841
9	0.13231	0.067877	0.02007	0.055919	0.1764

(Table 23) to be large and hence $\hat{\theta}_l$ to be dominant; in the case of trial 8, the random effects weighted mean $\hat{\theta}_l$ is given 74% of the weight in the empirical Bayes estimate. In contrast, trial 9 has a log odds ratio above 1, but a small variance which is well below the between-study variance (Table 23) and $\hat{\theta}_l$ only receives 22% of the weight in the empirical Bayes estimate. Owing to this small shrinkage factor the estimate is not pulled down as much as the estimate from trial 8 (Table 23). This results in the empirical Bayes estimate for trial 9 being larger than that for trial 8 (Figure 20).

Empirical Bayes estimates are a useful way of summarising meta-analysis data in the presence of heterogeneity. They allow the range of possible values of treatment effect to be seen, and hence may aid decisions regarding for which sort of patients the treatment is more (or less) effective. Furthermore, unlike the initial observed estimates, they do take account of all the information available. The use of different estimators of σ_B^2 may not have much impact on the empirical Bayes results. However, it may be worth carrying out a sensitivity analysis on a variety of values of σ_B^2 as none of the estimates of σ_B^2 are very precise and this imprecision is not taken into account in the empirical Bayes methods.

2.6.4 A review of the Bayesian approach to meta-analysis

Most Bayesian methodology in meta-analysis has been based on the use of hierarchical models and is closely related to the likelihood based random effects models. However, the approach is conceptually different and avoids the problem of having to assume that there is some population of studies from which the studies included in the meta-analysis are drawn at random. This is a great advantage of the Bayesian approach as the notion of a universal population is a common criticism levelled at the random effects meta-analysis models [33, 35]. DuMouchel and Harris [92] and Carlin [40] regard the studies in the meta-analysis as exchangeable. That is to say that they are viewed as each bearing on the same general question, with some dif-

ferences from study to study, but such that the differences cannot be anticipated a priori [40]. Skene and Wakefield [41] also use the exchangeability assumption when considering the analysis of multicentre trials with binary responses using Bayesian hierarchical models. They point out that care is required over this assumption as it is often the case that the investigators will have some prior idea of which centres are likely to be most effective with regards to treatment. In such cases they suggest the use of restricted exchangeability whereby the centres are split into groups by some characteristic, such as country, and the centres within each group are considered exchangeable. In general, the assumption of exchangeability means that an exchangeable prior distribution for the effects in the different studies may be assumed, so that the effects are independently and identically distributed conditional on the values of certain hyperparameters [40].

The model proposed by Carlin [40] is typical of the Bayesian approach to meta-analysis and is based on a normally distributed hierarchical model with three stages. As with the standard normally distributed random effects model it is assumed that the treatment effect in each trial has a normal distribution

$$\hat{\theta}_i \mid \theta_i, v_i \sim N(\theta_i, v_i) \tag{60}$$

and that the prior distribution [40, 42] is also normal and given by

$$\theta_i \mid \theta, \sigma_B^2 \sim N(\theta, \sigma_B^2) \tag{61}$$

Then for a Bayesian analysis, a third stage is required in which prior distributions are specified for θ and σ_B^2. Carlin [40] assumes a non-informative prior for both these parameters. The results for the overall mean treatment effect θ are obtained by writing down the likelihood and collecting terms so that the familiar standard random effects estimate and variance is obtained (equations (18) and (19)). The

posterior distribution for $\hat{\theta}_i$ conditional on θ and σ_B^2 has mean and variance given by

$$E(\theta_i \mid \hat{\theta}_1, ..., \hat{\theta}_k, \theta, \sigma_B^2) = (1 - B_i)\hat{\theta}_i + B_i\theta \tag{62}$$

$$Var(\theta_i \mid \hat{\theta}_1, ..., \hat{\theta}_k, \theta, \sigma_B^2) = (1 - B_i)v_i \tag{63}$$

where $B_i = v_i/(v_i + \sigma_B^2)$, the shrinkage factor defined in Section 2.6.2.

The posterior distribution of the θ_i conditional only on σ_B^2 may then be obtained by integrating the k independent normal distributions described by (62) and (63) over the posterior for θ described by (18) and (19) [40]. The resulting moments of this distribution are then

$$E(\theta_i \mid \hat{\theta}_1, ..., \hat{\theta}_k, \sigma_B^2) = (1 - B_i)\hat{\theta}_i + B_i\hat{\theta} \tag{64}$$

and

$$Var(\theta_i \mid \hat{\theta}_1, ..., \hat{\theta}_k, \sigma_B^2) = (1 - B_i)v_i + B_i^2 \frac{\sigma_B^2}{\sum_{i=1}^{k}(1 - B_i)} \tag{65}$$

Formula (64) is the same as that used in the empirical Bayes situation. However, in the empirical Bayes case, the point estimate of σ_B^2 calculated from the data would be used to obtain B_i. However, Carlin [40] indicates that there is a danger of underestimating the uncertainty in the resultant inferences since this prior variance can rarely be precisely estimated from the data, as was seen by the wide confidence intervals obtained for the estimate of between-study variance in Section 2.2.

A fully Bayesian solution, however, is computationally more complicated and involves the integration of each of the conditional distributions described by (18), (19), (62) and (63) over the posterior distribution of σ_B^2. Carlin, therefore, adopts

a Monte Carlo approach similar to that of Rubin [93] which provides approximate solutions to these equations.

Malec and Sedransk [42] start from the same basic model, but then use a more flexible prior for θ. This prior reflects the beliefs that there are subsets of θ_i such that the θ_i within each subset are similar but that there is uncertainty about the composition of such subsets. Gibbs sampling [94] may be used in order to obtain the empirical posterior distributions of the $\hat{\theta}_i$ which can then be used to obtain the desired unconditional posterior moments. Similarly, DuMouchel and Harris [92] used a series of hierarchical priors in a practical example to combine the results of cancer studies.

Skene and Wakefield [41] also used a hierarchical model in the analysis of multicentre trials, where they noted that the methodology is directly transferable to meta-analysis. Again a three stage model is required, but in contrast to Carlin, was based on the number of successes in each group given the underlying probabilities of success P_{ij}, $i = 1, ..., k$ and $j = 1, 2$ (2 treatment groups), as the outcome with which to work. The first stage assumes that the number of successes in each group in each centre follow an independent binomial distribution. Hence, the first stage of the model is a product of $2k$ binomial distributions. This is the model used by van Houwelingen et al. [45] in their bivariate random effects model which is an extension of the model described in Section 2.4.

At the second stage, a joint distribution for the P_{ij} is specified and Skene and Wakefield reparameterise the model so that

$$\lambda_i = log\frac{P_{i2}}{(1 - P_{i2})} \tag{66}$$

$$\theta_i = log\frac{P_{i1}}{(1-P_{i1})} - \lambda_i \tag{67}$$

Hence, λ_i is the logistic transform of the rate in the placebo group and θ_i is the corresponding log odds ratio for centre i. It is often reasonable to assume that $p(\lambda_i, \theta_i \mid \mu, \Sigma)$ is a bivariate normal distribution where $\mu = (\mu_\lambda, \mu_\theta)$ and $\Sigma = (\sigma_\lambda^2, \sigma_\theta^2, \rho)$. Then assuming exchangeability, which thus allows inferences on mean values even when the θ_i are different, and a prior distribution $p(\mu, \Sigma)$, the joint posterior density $p(\lambda, \theta, \mu, \Sigma \mid y)$ has the form

$$\prod_{i=1}^{k} \frac{e^{(\lambda_i+\theta_i)x_{i1}}}{(1+e^{\lambda_i+\theta_i})^{n_{i1}}} \frac{e^{\lambda_i x_{i2}}}{(1+e^{\lambda_i})^{n_{i2}}} \prod_{i=1}^{k} \mid \Sigma \mid^{-1/2} exp\{-\frac{1}{2}([\lambda_i, \theta_i]^T - \mu)^T \Sigma^{-1} ([\lambda_i, \theta_i]^T - \mu)\} p(\mu, \Sigma) \tag{68}$$

Evaluation of integrals involving the posterior density function (68) is required in order to characterise the posterior distribution and make useful inferences. This is computationally intensive, but may be done based on the repeated use of Gauss-Hermite rules over a Cartesian grid [95]. For higher dimensional problems a method based on the same iterative procedure in conjunction with importance sampling Monte Carlo integration [96]. The marginal density of μ_θ, given the data, then gives a summary of the difference between the two treatments and that of σ_θ^2 reflects the between-centre variability.

Eddy, Hasselblad and Shachter [43, 97], proposed a Bayesian method for meta-analysis which they named the confidence profile method. This method is very flexible and may be used to adjust for different types of trials, different treatments and also biases within the trials to be combined in a meta-analysis. The method requires prior distributions, likelihood functions and functions describing biases to be specified. Noninformative priors may be used, while a separate likelihood function is required for each type of trial, each type of treatment and each type of outcome measure.

The confidence profile method can also be extended to a hierarchical Bayes model where there is no single overall treatment effect. Solutions to the problems basically involve the combination of information according to Bayes formula, although the mathematics gets increasingly complicated with increasingly complex models. Hence, computer software has been developed along with this method [97].

Bayesian methods for meta-analysis avoid the conceptual problems of the standard random effects model and are very flexible in terms of the models which may be set up. However, they are computationally intensive and, until fairly recently, there have been technical difficulties arising in the calculation of the required marginal densities. However, with new advances in methodology such as the development of Gibbs sampling [98, 99], and the efficient implementations of the computer algorithms this has ceased to be such a problem. In comparison to analytic approximation techniques such as those proposed by Naylor and Smith [95] and Smith et al. [96] which require specialist software, the calculations for Gibbs sampling are much easier and hence have become increasingly popular.

The advantage of the hierarchical Bayes model, which is equivalent to the random effects model, is that by obtaining a distribution for σ_B^2 the variation in this estimate is being taken into account. This is not the case in empirical Bayes methods or the standard random effects methods, but it is the problem that the likelihood models address by using profile likelihoods to obtain confidence intervals for the parameters of interest (Sections 2.2 and 2.4).

Furthermore, Bayesian methods allow sensitivity studies to be undertaken to assess the robustness of the inferences to the choice of model, prior and error distributions [41]. For example, if one or two outlying studies are apparent, a heavy tailed density, such as the bivariate t-distribution could be used in place of the bivariate normal density in the Skene and Wakefield model at the second stage. Carlin [40] uses a weighted normal plot to assess the adequacy of the normal approximations, primarily

the normality of the random effects distribution. He also uses plots which show the dependence of the posterior mean on the value of σ_B^2 in a similar way to the sensitivity plots described in Section 2.1. Hence, there has perhaps been a greater emphasis on the checking of the assumptions underlying the models in a Bayesian context than in the classical meta-analysis framework. Considerable work in this thesis, however, is directed at applying model checking methods to standard meta-analysis methods (Section 2.1, Chapter 3 and Chapter 5).

2.7 Comparison of alternative methods of estimating the between-study variance

Four estimators of the between-study component of variance have been suggested so far; the D&L moment estimator (Section 1.7.1), the Maritz and Lwin moment estimator (Section 2.6.2) and the two maximum likelihood estimators (Sections 2.2 and 2.4). A further possible estimator of σ_B^2 is now considered, which has here been adapted to the meta-analysis case from a related situation. This new estimate of σ_B^2 is introduced and described in Section 2.7.1. It is then compared with the D&L estimate, to which it is closely related, in Section 2.7.2 using both simulated and practical examples.

2.7.1 Introduction

Matthews [100] considered the problem of analysing repeated or serial measurements of a continuous variable using summary measures. In this type of application, it is important to consider both the within-subject variation over the repeated measures as well as the between-subject variation. These two components of variance can be considered as equivalent to the within-study and between-study variances in a meta-analysis. The summary measure used for each subject is the regression coefficient

and the precision attached to each of these slopes may vary considerably between individuals. This leads to the consideration of a weighted analysis and hence the use of the following model [100]:

$$y_{ij} = \alpha + \beta_i t_{ij} + \varepsilon_{ij} \tag{69}$$

where y_{ij} is the j^{th} measurement, $j = 1, .., n_i$, on the i^{th} individual, $i = 1, .., k$, taken at time t_{ij} and $\varepsilon_{ij} \sim i.i.d.N(0, \sigma_i^2)$. Now assuming $\beta_1, ..., \beta_k$ are normally distributed with mean β_0 and variance σ_B^2, and may be estimated by the regression coefficients b_i, $i = 1, ..., k$,

$$b_i = \beta_i + \frac{\sum_{j=1}^{n_i} \varepsilon_{ij} t'_{ij}}{\sum_{j=1}^{n_i} t'^2_{ij}} \tag{70}$$

where t'_{ij} is t_{ij} measured from the mean of the times of measurements for subject i and $var(b_i) = \sigma_i^2 + \sigma_B^2$. Hence a weighted analysis can now be based on the weights $1/var(b_i)$. Assuming that $b_1, ..., b_k$ are independent, they have a marginal distribution of $N(\beta_0, \sigma_i^2 + \sigma_B^2)$ and hence an estimate of the between-subject variance is required. Matthews obtains such an estimate by equating the sample variance S_B^2 with its expectation. Rearranging the resulting expression produces an unbiased estimate of the between-subject variance which has the form

$$\hat{\sigma}^2_{BM} = S_B^2 - \frac{1}{k} \sum_{i=1}^{k} s_i^2 = S_B^2 - \overline{s_i^2} \tag{71}$$

where s_i^2 is the usual estimate of the variance of the slope. This expression (71) is directly applicable to the meta-analysis situation, where $S_B^2 = \frac{1}{k-1} \sum_{i=1}^{k} (\hat{\theta}_i - \overline{\theta}_i)$, the variance of the $\hat{\theta}_i$, where $\overline{\theta}_i$ is the simple unweighted mean of the $\hat{\theta}_i$ and $\overline{s_i^2} = \frac{1}{k} \sum_{i=1}^{k} v_i$. The Matthews estimate (71) is equivalent to the D&L estimate of σ_B^2, except that S_B^2 is used as the basis for the method of moments instead of Q. In fact, when all weights

are equal, then both these estimates of σ_B^2 are the same. As with the D&L estimator, the unbiased nature of $\hat{\sigma}_{BM}^2$ is lost as all negative values must be set to zero. The difference between the two estimators is that the D&L estimate takes into account the different precisions of the individual study estimates, whereas the Matthews estimator does not. Hence, intuitively, it would seem more sensible to use the D&L estimate. A comparison, therefore, of the Matthews and the D&L estimators of the between-study variance in a meta-analysis was undertaken.

2.7.2 Simulation results

Simulations were carried out in order to compare the Matthews estimate $\hat{\sigma}_{BM}^2$ with the closely related D&L estimate $\hat{\sigma}_B^2$ of the between-study variance in a meta-analysis context. The D&L estimate was used as a comparison, since it is constructed in a similar way to the Matthews estimate. Furthermore, both are simple to calculate, but may be inferior to maximum likelihood estimates. All calculations in the simulations were carried out using the true known values of v_i and the $\hat{\theta}_i$ generated from the model (see Section 3.3.1 for details of the computer simulation methods used) in order that any potential bias caused by the estimation of the v_i be avoided. In these examples the number of studies in the meta-analysis k was taken to be 10 and 1000 repetitions were executed each time. The three initial simulated examples were such that they represented cases with increasing amounts of variability in the precision of the individual study estimates of treatment effect. Hence, the v_i were taken to be equal, slightly different and severely different; the details of the actual values used are shown in Table 25. The sample mean and standard deviation of the 1000 simulated values of each estimator were obtained for both the distribution of the raw estimates of σ_B^2, that is including negative values, and the biased estimates, that is taking $\max\{0,\hat{\sigma}_B^2\}$ (Table 25). The true between-study variance σ_B^2 was set to 0.25 in all simulations.

Table 25: Simulation results comparing the performance of two moment estimators of between-study variance under varying conditions

Example	Estimator	Raw results		$\max\{0,\hat{\sigma}_B^2\}$	
		Mean of $\hat{\sigma}_B^2$	Standard deviation of $\hat{\sigma}_B^2$	Mean of $\hat{\sigma}_B^2$	Standard deviation of $\hat{\sigma}_B^2$
1	D&L	0.2544	0.6038	0.3679	0.4967
	Matthews	0.2544	0.6038	0.3679	0.4967
2	D&L	0.2518	0.2686	0.2647	0.2528
	Matthews	0.2481	0.3018	0.2679	0.2794
3	D&L	0.22561	1.9060	0.8459	1.3861
	Matthews	0.29526	3.1377	1.3359	2.3362

Key

The true values for all the simulations are:

Between-study variance σ_B^2=0.25

Overall treatment effect θ=5

Example 1: within-study variance v_i=1 for all i

Example 2: v_i=0.6,0.55,0.50,...,0.15

Example 3: v_i=10,9,8,...,1

In the example where all the v_i are equal, both estimators produce exactly the same results. This is due to the fact that in the equal weighting case, $\hat{\sigma}^2_{BM}$ reduces to $\hat{\sigma}^2_B$ and, since the true values of v_i were used, equal weighting is assured for each simulation. The result of equality can be shown algebraically, for if $v_i = v$ for all i which implies that $w_i = w$ for all i, then

$$\hat{\sigma}^2_B = \frac{w\sum_{i=1}^{k}(\hat{\theta}_i - \overline{\theta}_i) - (k-1)}{\left(kw - \frac{kw^2}{kw}\right)} \tag{72}$$

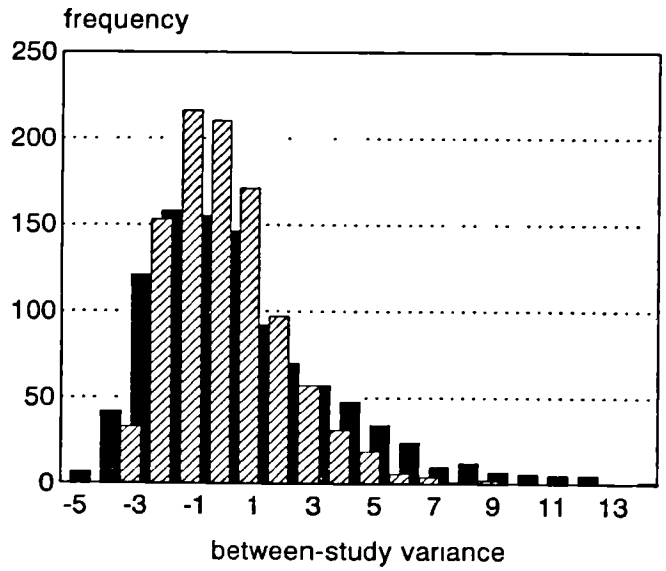
Now since $w = 1/v$, then equation (72) becomes

$$\frac{1}{(k-1)}\sum_{i=1}^{k}(\hat{\theta}_i - \overline{\theta}_i) - v = \hat{\sigma}^2_{BM} \tag{73}$$

As the v_i become increasingly more different, the performance of both the estimators deteriorates, with the estimates becoming more biased and less precise, but that of the Matthews estimator $\hat{\sigma}^2_{BM}$ does so to a greater extent (Figure 22). The two examples with unequal v_i illustrate that when the v_i are different the D&L estimator $\hat{\sigma}^2_B$ has a clear advantage over the Matthews estimator $\hat{\sigma}^2_{BM}$. In each of these examples the standard deviation of the D&L estimator is less than that of the Matthews estimator. Furthermore, when $\max\{0,\hat{\sigma}^2_B\}$ is taken, the fact that the Matthews estimator is more variable than the D&L estimator means that more values have to be set to zero. This leads to an increase in the sample mean of $\hat{\sigma}^2_{BM}$ leading to it becoming much larger than the true value of the between-study variance. Although the same thing happens with the D&L estimator, the smaller variability means fewer values being set to zero and hence a lesser effect on the mean (Figure 23). The standard deviation for both estimators becomes smaller when $\max\{0,\sigma^2_B\}$ is taken due to the restriction in the possible values.

The diuretics trials data were used in order to provide a practical comparison

Figure 22: Distributions of the DerSimonian and Laird estimator and the Matthews estimator of between-study variance from 1000 simulated meta-analyses for example 3 of Table 25

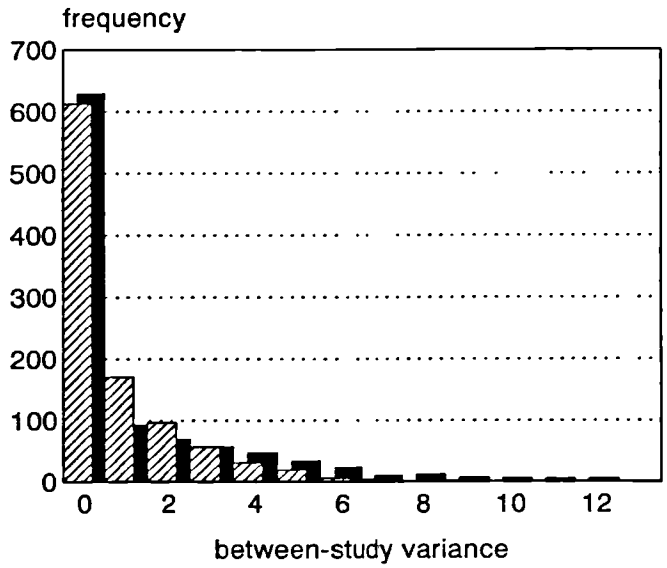


Key

▨=DerSimonian and Laird estimate of σ_B^2

■ =Matthews estimate of σ_B^2

Figure 23: Distributions of the DerSimonian and Laird estimator and the Matthews estimator of between-study variance using $max\{0, \hat{\sigma}_B^2\}$ from 1000 simulated meta-analyses for example 3 of Table 25



Key

▨ =DerSimonian and Laird estimate of σ_B^2

■ =Matthews estimate of σ_B^2

of the two estimators. The Matthews estimate $\hat{\sigma}^2_{BM}$ was calculated to be 0.51, which is very different to the D&L estimate which is only 0.23. Although this large difference in the estimate of the between-study variance did not have a great impact on the overall point estimate of treatment effect, the associated standard error was obviously much larger reflecting the extra uncertainty suggested by $\hat{\sigma}^2_{BM}$. In the light of the the simulations performed and since the precisions of the estimates in the diuretics trials do vary, the D&L estimate of the between-study variance is to be preferred, particularly since this estimate also closely agrees with the MLE.

The concern about whether an estimator should be based on a quantity which does not take into account the varying precision of the estimates has been found to be justified. The Matthews estimator of the between-study variance, at least in the examples simulated, has been found to be inferior to the D&L estimator in the meta-analysis context, and should not be used when precision of individual estimates vary considerably. The problem of estimating σ^2_B imprecisely is not helped by the new estimator, as the simulations have shown it to be less precise and more biased. In addition, the D&L estimator is as simple to calculate as the Matthews estimator and so is still to be preferred. However, the D&L estimator, it should be noted, did not perform well in all cases and was still biased. Hence, it may be that none of the moment estimators of the variance are very reliable and a likelihood estimate may be a better alternative.

2.8 Conclusion

The sensitivity plots presented in Section 2.1 are a useful way of investigating the robustness of the estimate of the overall treatment effect to changes in the value of σ^2_B. They provide information regarding the influence that an imprecise estimate of σ^2_B may have on the estimate of treatment effect. Since in practice neither a fixed effect or a random effects model is ideal, such a plot is valuable in assessing the

validity of the results obtained from the meta-analysis. Furthermore, the plots may be used to help determine whether the use of a likelihood meta-analysis model is required.

Likelihood models, such as that proposed in Section 2.2, produce confidence intervals for θ which are wider than those obtained from the standard random effects methodology, although this increase may be very negligible in many cases. Thus using a likelihood approach may lead to a more cautious interpretation of the results. The increase in the width of the confidence interval obtained from the likelihood model in Section 2.2 is due to the fact that such a model overcomes one of the problems with the standard random effects analysis, that is it allows for the estimation of σ_B^2 in the calculation of the confidence interval for θ. Furthermore, unlike the standard random effects method, a likelihood approach allows a confidence interval for σ_B^2 to be obtained. The examples presented in Section 2.3 suggest that caution is required when interpreting results from the standard random effects methods, particularly when the meta-analysis is based on only a small number of trials. However, when the numbers of trials in a meta-analysis is large, then the standard method is often adequate. The quadratic approximation to the likelihood method outlined in Section 2.3.4 is of little practical use since the confidence intervals tend to be too narrow, and furthermore, computationally it is no simpler than the method based on the profile likelihood. In fact, more work is required when a transformation of the parameter of interest is necessary to obtain reasonable quadratic approximations, as was the case in the example presented. Hence, if a likelihood approach is to be pursued the confidence interval should be obtained directly from the relevant profile log-likelihoods rather than by quadratic approximation.

The Mantel-Haenszel-type random effects likelihood procedure for binary data described in Section 2.4 offers further theoretical improvements over the marginal likelihood method of Section 2.2. This is because such a model is based on the exact

distribution of the 2×2 contingency table for each trial i, $i = 1, ..., k$, and thus the estimation of the weights w_i is taken into account. This implies that the confidence interval for θ will, if anything, be even wider than that obtained from the marginal likelihood model. In the examples considered (Section 2.4.3 and Section 2.5.3) both likelihood models produced very comparable results and no clear practical advantage was gained from the use of the Mantel-Haenszel-type procedure. The two methods produced very similar confidence intervals for θ, even in cases where there were zero event rates (Section 2.5.3), although, in general, the full likelihood approach should probably be preferred when there are small numbers of observations in all or some of the trials since it is based on the exact distribution as opposed to an asymptotic approximation. Furthermore, the Mantel-Haenszel-type model may be adapted to model the random effects using nonparametric methods in cases where the assumption of normality is unreasonable. The marginal likelihood method does, however, have certain advantages over the Mantel-Haenszel-type method. It is more flexible in that it can be used to analyse continuous measures as well as binary. Such a model based on continuous outcome measures is similar to an analysis based on a mixed model which may be implemented in the computer software package SAS [101]. Furthermore, for binary outcome measures the marginal likelihood method only requires $\hat{\theta}_i$ and v_i for each trial, whereas the Mantel-Haenszel-type procedure requires the full 2×2 contingency table for each trial, which may not always be readily available in the published literature. From the evidence provided by the examples in this chapter, the choice between the two likelihood methods may depend more on the ease with which the method can be implemented using a computer. A computer program is available from van Houwelingen et al. [45], written in Gauss, using the EM algorithm to obtain the likelihood solutions for the Mantel-Haenszel-type model. The work relating to the marginal likelihood model was carried out using GLIM where a macro containing a simple cyclical iteration procedure was used to obtain the MLEs, and also in Splus where standard Splus functions were used.

A Bayesian approach to meta-analysis was briefly introduced in Section 2.6, as the ideas are related to the likelihood approach previously considered. Empirical Bayes estimates (Sections 2.6.1–2.6.3) were shown to be a useful alternative to a single overall estimate of treatment effect for summarising the results from a meta-analysis in the presence of heterogeneity. It was found that the use of alternative estimators of σ_B^2 in the calculation of the empirical Bayes estimates $\tilde{\theta}_i$ is unlikely to have a great deal of influence on the results. However, the robustness of each $\tilde{\theta}_i$ to different estimates of σ_B^2 may be worth checking, particularly if $\hat{\sigma}_B^2$ is imprecise owing to the meta-analysis being based on a small number of trials.

The hierarchical Bayes model (Section 2.6.4) is similar to the random effects likelihood model, although by using the concept of exchangeability the controversial assumption that the trials included in a meta-analysis are a random sample from some global population of trials is avoided. Furthermore, the choice of the model, prior and error distributions is flexible in the Bayesian framework. However, methods tend to be computationally intensive and are perhaps conceptually more difficult for the non-statistician than the standard meta-analysis methods. The existing Bayesian meta-analysis literature also describes analyses to assess the robustness of the results to various modelling assumptions. Plots similar to those of Section 2.1 have been used in the Bayesian context, as have normal plots assessing the adequacy of the distributional assumptions similar to those proposed in the next chapter of this thesis. The emphasis on sensitivity analyses shown in the Bayesian context is valuable and the ideas could be usefully translated to the classical framework. Hence, although it is acknowledged that Bayesian methodology has much to offer meta-analysis it is not pursued any further here. The focus of the thesis is to consider the more standard and accessible approaches which are currently widely used in practice.

Various different estimators of the between-study variance σ_B^2 have been proposed. However, none are very precise, particularly when k is small, and all become

biased as negative values must be set to zero. Section 2.7 considered an alternative estimate of the between-study variance which was novelly applied to the meta-analysis context. The estimate, based on that described by Matthews [100] when considering the problem of the analysis of repeated and serial measurements of a continuous variable using summary measures, was adapted to the meta-analysis situation. The estimator was found to be unsuitable for application to meta-analysis, and certainly performed considerably worse than the D&L moment estimator of σ_B^2 in the simulation examples considered. It was less precise and more biased, particularly in examples where the weight was unevenly distributed. However, these examples also revealed that the D&L estimator was far from satisfactory in these unevenly weighted situations too.

This chapter has shown that the estimation of the between-study variance in a meta-analysis is problematic. Estimates tend to be imprecise, while allowing for the estimation of σ_B^2 may affect the estimate of the overall treatment effect. However, in many practical situations the use of the standard random effects method, which incorrectly assumes that σ_B^2 is known, will produce reliable results in the presence of heterogeneity as results can often be robust to changes in the value of σ_B^2. However, the implications of varying σ_B^2 and allowing for its estimation by means of a likelihood model should always be carefully considered. Furthermore, methods based on exact distribution theory should be considered for use with sparse binary data.

3 Checking Distributional Assumptions

To obtain confidence intervals for the fixed effect estimate of treatment effect the assumption that the individual study estimates are normally distributed with mean θ and variance v_i must be made. Additionally, in the random effects model for meta-analysis, it is usual to make the assumption that as well as the individual study estimates having a normal distribution the random effects have a normal distribution too. Many of the results and methods in Chapters 1 and 2 are based on these distributional assumptions. Section 3.1 describes the use of normal probability plots to check the distributional assumptions of both the fixed effect and the random effects model, while the issue of testing for normality is addressed in Section 3.2. Section 3.3 investigates the performances of the plots and tests using simulation techniques and Section 3.4 provides some practical examples of the use of the methods. The chapter is completed with a discussion in Section 3.5.

3.1 Normal Plots

The fixed effect version of the normal plot will firstly be described in Section 3.1.1, followed by the corresponding random effects plot in Section 3.1.2.

3.1.1 Fixed effect plot

The contribution that study i makes to the test statistic for heterogeneity Q can be written as $q_i = (\hat{\theta}_i - \hat{\theta}_f)/\sqrt{v_i}$, where $\sum_{i=1}^{k} q_i^2 = Q$. Under a normally distributed fixed effect model, the $(\hat{\theta}_i - \theta)/\sqrt{v_i}$ will have a standard normal distribution. Hence, the distribution of q_i, where θ is replaced by $\hat{\theta}$, will be approximately standard normal, provided that k is large. The q_i are ordered such that $q_{(1)} \leq q_{(2)} \leq ... \leq q_{(k)}$, and then a normal plot, or 'q-q plot', is a display of $q_{(i)}$ against $\Phi^{-1}(F_k(q))$ where $\Phi(q)$

and $F_k(q)$ are the standard normal and empirical cumulative distributional functions (cdf), that is

$$F_k(q) = \sum_{i=1}^{k} I(q - q_{(i)})/k \tag{74}$$

where $I(x)$=1 for $x \geq 0$ and 0 otherwise. It should be noted that in practice adjustments must be made at the endpoints, since $\Phi^{-1}(F_k(q))$ cannot be calculated for $q \leq q_1$ or $q \geq q_k$. Blom scores [102] which are written as $F_k(q_{(i)}) = (i-3/8)/(k+1/4)$ may be used for this purpose. A q-q plot of $q_{(i)}$ will produce, approximately, a straight line through the origin with unit gradient if the specified distributional assumptions hold. Hence, the plot can be used to provide a visual inspection of the validity of the normality assumption that $\hat{\theta}_i \sim N(\theta, v_i)$ in a fixed effect model.

3.1.2 Random effects plot

A plot, corresponding to that described in the Section 3.1.1 for the fixed effect model, may be produced to check the normality assumptions underlying the random effects model, that is $\hat{\theta}_i \sim N(\theta_i, v_i)$ and $\theta_i \sim N(\theta, \sigma_B^2)$. A similar statistic to q_i can be used, except that the fixed effect estimate $\hat{\theta}_f$ is replaced by the random effects estimate $\hat{\theta}_r$ and the additional component of variance σ_B^2 is incorporated. The statistic used is therefore $q_i^* = (\hat{\theta}_i - \hat{\theta}_r)/\sqrt{(v_i + \hat{\sigma}_B^2)}$. Under the normality assumptions, the q_i^* follow an approximate standard normal distribution. Therefore, under a normally distributed random effects model, a q-q plot of the ordered q_i^* will produce, approximately, a straight line through the origin with unit gradient.

Since $\hat{\sigma}_B^2$ is always greater than or equal to zero, then $\sqrt{(v_i + \hat{\sigma}_B^2)}$, the denominator of q_i^*, is always greater than or equal to $\sqrt{v_i}$, the denominator of q_i. Hence, because $\hat{\theta}_f$ is usually very close in value to $\hat{\theta}_r$, $|q_i^*|$ will be less than or equal to $|q_i|$.

This means that each point will be shrunk towards zero in transferring from the fixed effect plot to the random effects plot.

In addition, if $v_i = v$ for all i, then under the normally distributed random effects model, q_i will have an approximate normal distribution with a mean of zero and a variance given by $(v + \hat{\sigma}_B^2)/v$. Hence, using q_i when the data actually follow the normally distributed random effects model will produce a plot which is a straight line going through the origin and which has a gradient which is steeper than one. In this situation, where all within-study variances are equal, each point is pulled in towards zero by the same proportionate amount when comparing the random effects plot to the fixed effect plot. In fact, q_i^* can be written as cq_i, where c is a constant which takes the value $\sqrt{v/(v + \hat{\sigma}_B^2)}$. However, it becomes much more difficult to predict how the points will be transformed on the random effects plot, in comparison to the fixed effect plot, when the variances are different. The proportion by which a point moves then depends on the ratio of each individual within-study variance to the between-study variance. A point with a small within-study variance v_i in relation to between-study variance $\hat{\sigma}_B^2$ will change by a greater extent in proportionate terms than a point with a larger within-study variance, since q_i^* is now equal to $(\sqrt{v_i/(v_i + \hat{\sigma}_B^2)})q_i$. However, in absolute terms a trial which is in the tails of the distribution of q_i, and thus with large $(\hat{\theta}_i - \hat{\theta}_f)^2$, will tend to change more than one in the centre of the distribution. The points may also have different normal score values in the two plots meaning, therefore, that the ordering of the points changes.

Examples of both types of normal plots are presented later in this chapter.

3.2 Testing for Normality

Since it may be difficult to judge from an informal visual inspection of the q-q plots whether the data is compatible with a standard normal distribution, a test of the

normality would be of use. The following sections (Section 3.2.1 and 3.2.2) describe two possible tests for this purpose.

3.2.1 The Shapiro-Francia W' test

The Shapiro-Francia W' test [103] is a powerful test of departure from normality [104] which is straightforward to carry out. Using the ordered q_i, the test is of the hypothesis that the sample of interest is from a normal distribution with an unknown mean μ and an unknown variance σ^2. These unknown parameters may be estimated as follows,

$$\hat{\mu} = \frac{1}{k}\sum_{i=1}^{k} q_{(i)} \tag{75}$$

and

$$\hat{\sigma} = \frac{\boldsymbol{m}^T\boldsymbol{q}}{\boldsymbol{m}^T\boldsymbol{m}} \tag{76}$$

where $\boldsymbol{q}$ is the vector of the $q_{(i)}$ and $\boldsymbol{m}$ is the expectation vector of the order statistics of a sample of standard normal random variables. The Shapiro-Francia W' test statistic is then defined by,

$$W' = (\sum_{i=1}^{k} a_{(i)}q_{(i)})^2 / \sum_{i=1}^{k}(q_{(i)} - \overline{q})^2 \tag{77}$$

where $\boldsymbol{a} = (\boldsymbol{m}^T\boldsymbol{m})^{1/2}\boldsymbol{m}$. In practice, $\boldsymbol{m}$ may be approximated by the Blom scores $\tilde{\boldsymbol{m}}$, where

$$\tilde{m}_{(i)} = \Phi^{-1}\{(i - 3/8)/(k + 1/4)\} \tag{78}$$

Royston [104] provides a method for transformation of the null distribution of W' to

normality so that the p-value for the test may be obtained.

It should be noted that $\sum_{i=1}^{k}(q_{(i)} - \bar{q})^2 = S^2$ is the usual unbiased estimate of $(k-1)\sigma^2$ where σ^2 is the variance of the $q_{(i)}$. When the sample comes from a normal distribution, then both the numerator of (77), $(\sum_{i=1}^{k} a_{(i)}q_{(i)})^2$, and the denominator S^2, are, apart from a constant, estimating the same quantity, namely σ^2 [105]. Hence, W' will be approximately equal to $(k-1)^{-1}$. If the sample comes from a non-normal population then these two quantities are not, in general, estimating the same thing and so W' will not be equal to $(k-1)^{-1}$. Since this test is only concerned with whether the sample is from a normal distribution it actually tests only the linearity of the plot. However, the hypothesis of interest here is that the observed sample is from a standard normal distribution. Hence, the interest lies in the deviation of the points from the line of identity, not simply in the deviation from linearity. This means that a more useful test would be one which looked at the gradient of the plot as well as the linearity.

An estimate of the slope of the plot may be obtained using the formula given in (76) and hence this will give some indication of whether the slope is consistent with unity. This implies that a calculation of the slope of the regression line of the ordered values together with the test for linearity may indicate whether the assumption of standard normality is reasonable. However, a single test of standard normality $N(0,1)$ would be preferable.

3.2.2 The Anderson-Darling A^2 test

Empirical distribution function (EDF) statistics for goodness-of-fit offer the possibility of an improvement over the Shapiro-Francia W' test as they do allow for the complete specification of the distribution of the null hypothesis. The most commonly used EDF goodness-of-fit test is the Kolmogorov test, which looks at the maximum

distance between the hypothesised distribution and the empirical distribution. However, due to the fact that it is only considering a maximum difference, this test is not the most powerful EDF test. From a comparison study of 5 EDF tests, Stephens [106] suggested that it was always worthwhile considering the Cramer-von Mises statistic W^2, the Watson statistic U^2 and the Anderson-Darling statistic A^2. Sinclair and Spurr [107] indicate that the Anderson-Darling test is designed to be more sensitive to discrepancies between $F_k(q)$ (74) and $F(q) = \Phi(q)$, the standard normal function, in the tails of the distribution, a feature which would be useful for the particular case of meta-analysis under consideration. Hence, the Anderson-Darling statistic is the EDF test considered here. Letting $F(q_{(i)}) = z_{(i)}$, the Anderson-Darling statistic is given by

$$A^2 = -\frac{\sum_{i=1}^{k}(2i-1)[ln(z_{(i)}) + ln(1 - z_{(k+1-i)})]}{k} - k \tag{79}$$

This statistic is based on the idea that the distribution of $F(q_{(i)})$ is symmetric under the null hypothesis, that is $z_{(i)} = z_{(k+1-i)}$. Hence, under the null hypothesis

$$A^2 = -\frac{\{\sum_{i=1}^{k}(2i-1)ln(z_{(i)}^2)\}}{k} - k \tag{80}$$

Furthermore, under the null hypothesis $z_{(i)}$ will take the value of the midpoint of $F_k(q_{(i-1)})$ and $F_k(q_{(i)})$ so that $z_{(i)} = (\frac{i}{k} + \frac{i-1}{k})/2 = \frac{2i-1}{2k}$. Hence (80) can be written as

$$A^2 = -\frac{2}{k}\left\{ \sum_{i=1}^{k}(2i-1)ln(\tfrac{2i-1}{2k}) \right\} - k \tag{81}$$

Then as k becomes reasonably large, it can be shown that $-\frac{2}{k}\left\{ \sum_{i=1}^{k}(2i-1)ln(\frac{2i-1}{2k}) \right\}$ is approximately equal to k. This means that under the null hypothesis $A^2 \simeq k - k = 0$. Therefore as $F_k(q)$ becomes more different from $F(q)$, A^2 will get larger and it is this characteristic that provides the basis for the test.

Different versions of the Anderson-Darling test for normality are required for different circumstances and the choice of test depends on what facts are specified about the null distribution of the parameter of interest. There are four different variants of the null hypothesis of normality:

(1)Mean (μ) and variance (σ^2) known
(2)Mean known and variance unknown
(3)Mean unknown and variance known
(4)Mean and variance unknown

Number (1) is the version of interest here, as in the present problem both the mean and the variance of the null distribution of the q_i are assumed to be known, although the estimation of $\hat{\theta}_f$ or $\hat{\theta}_r$ means that the results are only approximate. Now each $z_{(i)}$ is calculated by standardisation where $z_{(i)} = (q_{(i)} - \mu)/\sigma$, but in the situation under consideration, however, μ=0 and σ=1 and so $z_{(i)} = q_{(i)}$. In the other three cases where there are unknown parameters, estimates of μ and σ^2 may be obtained from the observed $q_{(i)}$, and then $z_{(i)}$ can be calculated using $\hat{\mu}$ and $\hat{\sigma}^2$ as required. The parameters μ and σ may be estimated by $\hat{\mu} = \sum_{i=1}^{k} q_{(i)}/k$ and $\hat{\sigma^2} = \sum_{i=1}^{k}(q_{(i)} - \hat{\mu})^2/(k-1)$ or $\sum_{i=1}^{k}(q_{(i)} - \mu)^2/k$ depending on what has been specified about the null distribution. The resulting $z_{(i)}$ are then used to calculate the test statistic A^2. Stephens [106] provides a separate table of critical values for each of the four different cases described above. For the case where both the mean and the variance are estimated a modification is made to A^2 before it is looked up in a table, namely $A^{2*} = A^2(1+4/k-25/k^2)$.

After the necessary transformation of $q_{(i)}$ to a standard normal distribution $z_{(i)}$, the hypothesised cumulative distribution function $F(q_{(i)})$ is completely specified and the $z_{(i)}$ should be uniformly distributed between 0 and 1. If the mean of the sample is different to that specified then the points will tend to move towards 0 or 1.

If the variance is different to that specified then the points will tend to move towards each end or towards 0.5 [106].

The performance of both the Shapiro-Francia W' and the Anderson-Darling A^2 tests are compared in Section 3.3 under different null and alternative hypotheses.

3.3 Simulation Studies

In order to compare the two tests for normality under different conditions, a series of simulations were carried out using the methods described in Section 3.3.1. The shape of both fixed effect q-q plots and random effects q-q plots were considered for examples of data generated under different models in Section 3.3.2. The simulations also presented an opportunity to investigate whether the estimation of θ in the calculation of q_i, $i = 1, ..., k$, has an effect on the results of either test being compared. This question is considered in Sections 3.3.3 and 3.3.4 in order that a valid and appropriate test be identified (Section 3.3.5). The main aspect of the investigation, addressed in Sections 3.3.6 and 3.3.7, is to see how well the Shapiro-Francia test and the Anderson-Darling test are able to distinguish between correctly and incorrectly specified meta-analysis models. Hence, the power of the two tests are investigated and compared under various conditions.

3.3.1 Description of simulation methods

The computer program used to carry out the simulations, written in FORTRAN, was designed to produce data points sampled from either a normally distributed fixed effect model or a normally distributed random effects model. The routines for generating the data were used for the investigations in Section 2.7 and Chapters 4 and 5, as well as for the work in the current chapter. Two different situations were considered which involved the simulation of different models. Firstly, the v_i, $i =$

$1, ..., k$ were assumed known, that is the true values from which the data are generated are used in any subsequent calculations. Such examples involve the generation of values of $\hat{\theta}_i$ such that

$$\hat{\theta}_i \sim N(\theta_i, v_i) \tag{82}$$

where each θ_i is generated from the distribution

$$\theta_i \sim N(\theta, \sigma_B^2) \tag{83}$$

These values of $\hat{\theta}_i$, together with the known v_i, are then used in the calculation of $\hat{\theta}_f$, $\hat{\sigma}_B^2$ and $\hat{\theta}_r$ by standard meta-analysis methods. The data, both θ_i and $\hat{\theta}_i$, were obtained using a random number generator, followed by a transformation to normality. In the case where the known parameters are used, it is not necessary to generate individual data points within each study. Hence, the results from such simulations may apply to either a binary or a continuous outcome measure. However, the conclusions drawn cannot be directly applied to a practical situation, since the estimation of the parameters may affect the results.

In order to investigate what happens in practice where v_i as well as θ_i must be estimated from the data within each study, individual data points must be generated in the simulations. The results presented in this Section are based on a continuous outcome measure, as this data is less problematic to work with than simulated binary data, and it also relates to the major practical example used to illustrate the ideas in this chapter (Section 3.4.1). For treatment group j (j=1,2) in trial i, an individual observation y_{ijl}, l=1,...,n_i, could be generated and the difference in means between groups calculated. However, in order to simplify the simulations, each data point y_{il}, l=1,...,n_i and $i = 1, ..., k$, is generated using

$$y_{il} \sim N(\theta_i, \sigma_i^2) \tag{84}$$

so that $\hat{\theta}_i \sim N(\theta_i, v_i)$, where $v_i = \hat{\sigma}_i^2/n_i$ and $\hat{\sigma}_i^2 = \sum_{l=1}^{k}(y_{il} - \overline{y}_{i.})^2/(n_i - 1)$, and as before $\theta_i \sim N(\theta, \sigma_B^2)$. The $\hat{\theta}_i$ and v_i, used in the subsequent calculation of $\hat{\theta}_f$, $\hat{\sigma}_B^2$ and $\hat{\theta}_r$ are obtained from the data y_{il} within each study using the standard techniques. For this, as well as the previous situation, data following a fixed effect model may also be produced from the same generating procedure. If $\hat{\theta}_i$ is from a fixed effect model, σ_B^2 may be set to zero and each θ_i is equal to θ and so $\hat{\theta}_i \sim N(\theta, v_i)$.

For the particular issue being investigated in this chapter, subroutines were written for the purpose of calculating $q_{(i)}$, $q_{(i)}^*$ and the Shapiro-Francia and Anderson-Darling test statistics, and 1000 data sets were generated for each example to obtain the required results. The number of observations n_i in each trial was set to be 50 for all the simulations in this section.

3.3.2 Examples of the plots

Before pursuing the main simulations, examples of the type of plots, both fixed effect (Section 3.1.1) and random effects (Section 3.1.2), that may be obtained from different types of data are presented. A single simulated example was taken from a selection of models where k was set at 50 and where all parameters were assumed known. This is simply as an introduction to provide a guide to the sort of plots that may be expected from different types of data. The number of points available was deliberately chosen to be large to enable the shape of the plots to be seen clearly. Furthermore, at this stage the parameters were assumed to be known, and the models generated using (82) and (83), as there is then no problem with the validity of the assumptions regarding the distribution of the $q_{(i)}$ or the validity of the tests. Initially, data sets which conform to the two standard meta-analysis models are considered, being

Model (1) Fixed effect model with $v_i = v{-}0.1$ for $i = 1, ..., k$, k=50

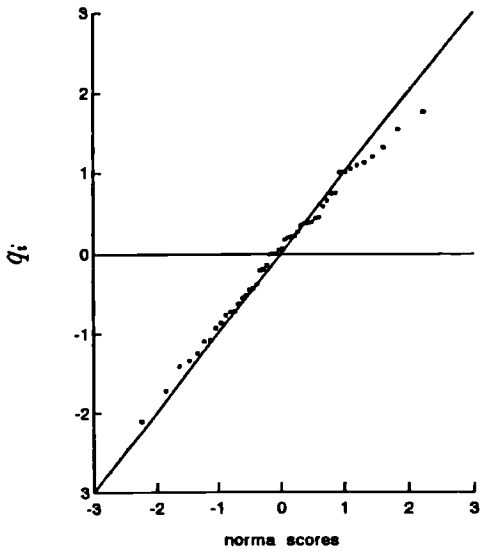
Model (2) Random effects model with σ_B^2=0.5 and $v_i = v$=0.1 for $i = 1, ..., k$, k=50

Model (3) Random effects model with σ_B^2=0.5 and different v_i,

$v_1 - v_5$=0.05, $v_6 - v_{10}$=0.10,...,$v_{45} - v_{50} = 0.5$

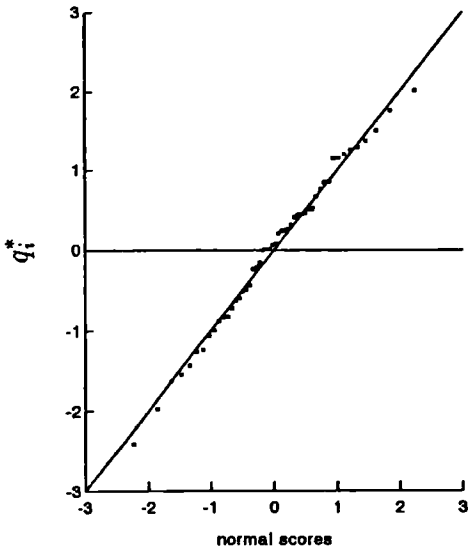
When the data is sampled from a homogeneous normal distribution with equal within-study variances (Model (1)), both the fixed effect plot (Figure 24) and the random effects plot (Figure 25) follow the line of identity with any deviations being compatible with chance. The random effects plot is very similar, although actually not identical, to the fixed effect plot. The plots would be identical if $\hat{\sigma}_B^2$=0. However, in the example $\hat{\sigma}_B^2$ must be slightly greater than 0 meaning that $q^*_{(i)} = cq_{(i)}$ for all i where $c = \sqrt{v/(v + \sigma_B^2)}$ (Section 3.1.2) is marginally less than 1.

Figure 25: Random effects normal plot of q_i from a normally distributed fixed effect model (k=50) compared with the $N(0,1)$ line



correlation=0.996

Figure 24: Fixed effect normal plot of q_i^* from a normally distributed fixed effect model (k=50) compared with the $N(0,1)$ line



correlation=0.996

The test results for both the Shapiro Francia W' test and the Anderson-Darling A^2 test in this example indicate that there is no evidence against normality in either plot (Table 26). It is noticeable that the value of the statistic W' is the same for both plots. This will in fact always be the case in examples where v_i=v for all i since the within-study variances are all equal, and $\hat{\theta}_f$ is approximately equal to $\hat{\theta}_r$, and $q^*_{(i)}$ has been shown to be equal to $cq_{(i)}$. Furthermore, since $\tilde{m}_{(i)}$ is the same for both $q_{(i)}$ and $q^*_{(i)}$ in the Shapiro-Francia test, then $a_{(i)}$ is the same for both the fixed effect and the random effects version of the test. For the random effects test using the $q^*_{(i)}$, therefore,

$$W' = (\sum_{i-1}^{k} a_{(i)} cq_{(i)})^2 / \sum_{i=1}^{k} (cq_{(i)} - c\bar{q})^2 = c^2 (\sum_{i-1}^{k} a_{(i)} q_{(i)})^2 / c^2 \sum_{i-1}^{k} (q_{(i)} - \bar{q})^2 \qquad (85)$$

and hence it can be seen that the c's cancel out, thus giving the test for the $q_{(i)}$. Hence, the test statistic W' is equivalent in the two situations, and only the linearity

of the plot is being tested.

Table 26: Results of two tests for normality for the simulated examples from models (1)–(5) shown in Figures 24–24

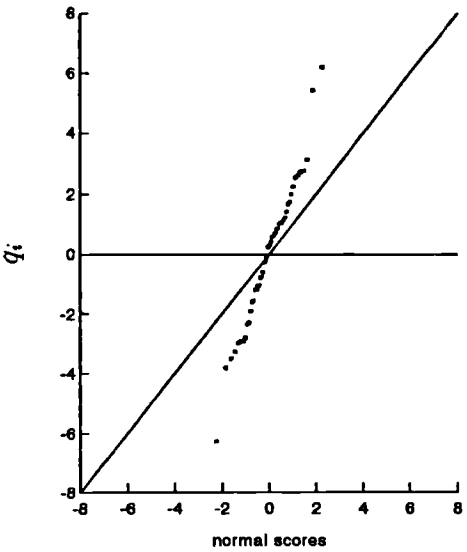
Model	Type of plot	Shapiro-Francia W' statistic	p-value for W'	Anderson-Darling A^2 statistic	p-value for A^2
(1)	Fixed effect	0.992	0.95	0.331	>0.10
	Random effects	0.992	0.95	0.179	>0.10
(2)	Fixed effect	0.977	0.36	14.663	<0.01
	Random effects	0.977	0.36	0.315	>0.10
(3)	Fixed effect	0.733	<0.001	8.375	<0.01
	Random effects	0.965	0.13	1.369	<0.025
* (4)	Fixed effect	0.942	0.02	1.199	<0.05
	Random effects	0.942	0.02	1.250	<0.05
* (5)	Fixed effect	0.921	0.004	1.462	<0.025
	Random effects	0.953	0.045	1.186	<0.05

* see p 163

If the data follow a random effects model with each study estimate having the same variance v (Model (2)), the fixed effect plot (Figure 26) produces a straight line through the origin with a gradient steeper than 1. This gradient will actually be approximately equal to $(v + \sigma_B^2)/v$ ($(v + \sigma_B^2)/v$=6 here) and the gradient of the regression line can be estimated using equation (76). The random effects plot (Figure 27) suggests that the model fits the data with only chance deviations from the line of identity. In this instance, it can be seen from the random effects plot that the distribution of the $q^*_{(i)}$ is approximately standard normal and from the fixed effect plot that the distribution of the $q_{(i)}$ is not. The estimated gradient of the line , using (76), on the fixed effect plot is 5.6 (compared to the theoretical value of 6), which is clearly greater than one, while the estimated gradient of that on the random effects

plot is approximately one.

Figure 26: Fixed effect normal plot of q_i from a normally distributed random effects model with equal within-study variances (k=50) compared with the $N(0,1)$ line

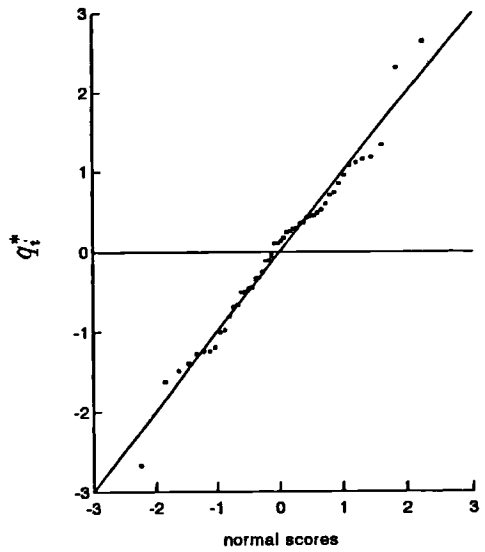


correlation=0.988

The Shapiro-Francia test applied to both the fixed effect and random effects plots find no evidence against normality (Table 26), and the test is again necessarily the same in both cases. The same argument as above (85) shows that this will always be the case when all the within-study variances are equal for random effects models as well as fixed effect models. The Anderson-Darling statistic is, however, able to detect that the $q_{(i)}$ are not standard normal and consequently does produce a significant result (Table 26). Furthermore, the result of A^2 using the random effects $q^*_{(i)}$ is non-significant as expected.

When the within-study variances are different under a random effects model (Model (3)), the fixed effect plot (Figure 28) still tends to produce a line with a gradient steeper than unity going through the origin. However, this line tends to be curved as opposed to straight, with the gradient becoming steeper towards the tails as v_i becomes larger for outlying points. The random effects plot (Figure 29)

Figure 27: Random effects normal plot of q_i^* from a normally distributed random effects model with equal within-study variances (k=50) compared with the $N(0,1)$ line

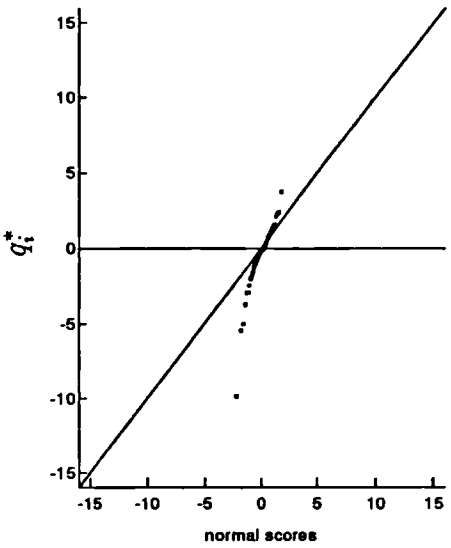


correlation=0.988

indicates that a normally distributed random effects model may reasonable. The Shapiro-Francia test correctly rejects normality of the fixed effect q_i, but finds the data consistent with a random effects model using the random effects q_i^*, thus correctly identifying the true model (Table 26). The Anderson-Darling test rejects the fixed effect model correctly, but the result for the test of the random effects is a false positive for this example.

As well as helping to decide between the two standard normally distributed models, these plots can also be useful in identifying sets of data which do not conform to either. For example, a data set which is a mixture of two distributions where most of the studies are homogeneous but where a few follow a random effects model, may be identified. Hence, two such examples were considered,

Figure 28: Fixed effect normal plot of q_i from a normally distributed random effects model with unequal within-study variances (k=50) compared with the $N(0,1)$ line



correlation=0.857

Model (4) Data a mixture of fixed effect and random effects models

Moderate between-study variance σ_B^2=0.5, v_i=v=0.1 for $i = 1, ..., k$ k=50

Number of points from a random effects model is 10

Model (5) Data a mixture of fixed effect and random effects models

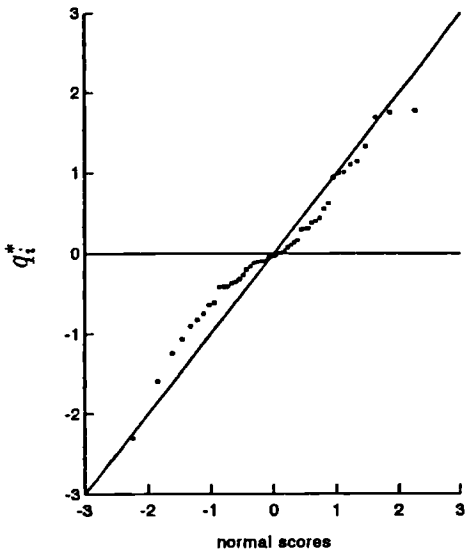
Moderate between-study variance σ_B^2=0.5, different v_i,

v_i varied from 0.05 to 0.5 as for Model (3), k=50

Number of points from a random effects model is 10

For a situation where 40 of the studies follow a normally distributed fixed effect model and the remaining 10 follow a normally distributed random effects model and where the v_i are all equal (Model (4)), the fixed effect plot (Figure 30) produces a display where the majority of the points follow the line of identity, but where some clear outliers, which fall well away from the line, can be observed in the tails of the

Figure 29: Random effects normal plot of q_i^* from a normally distributed random effects model with unequal within-study variances (k=50) compared with the $N(0,1)$ line

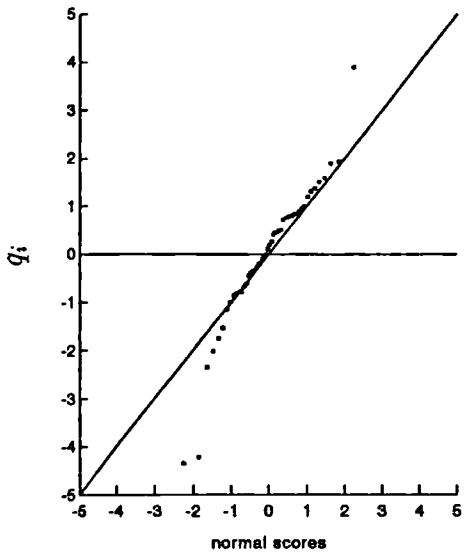


correlation=0.982

distribution. The corresponding random effects plot (Figure 31) appears to be more linear since the outlying points are pulled in towards the line of identity. However, the majority of the points, although still forming a relatively straight line have a gradient which is slightly less than 1. In this example, where the study variances are all equal, the Shapiro-Francia test will again produce the same result for both plots, as the two differ only in scale. Both A^2 and W' correctly identify this data set as being neither a fixed effect or a random effects model (Table 26). W' is specifically a test of linearity and hence is able to detect the deviations in the tails, while A^2 is also sensitive to deviations in the tails.

This evidence suggests that if a random effects plot is obtained where the gradient of most of the data is only slightly less than one on visual inspection, then caution over the normality of the plot should be expressed. Hence, a test for normality in this instance is helpful in detecting the non-normality because naive interpretation

Figure 30: Fixed effect normal plot of q_i from a data set which is a mixture of two distributions with equal within-study variances (k=50) compared with the $N(0,1)$ line

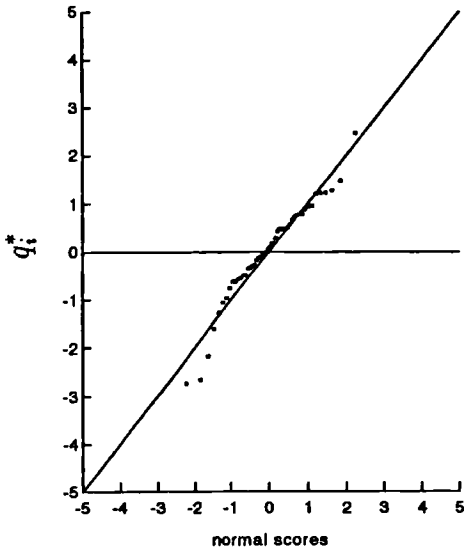


correlation=0.970

of the plots may be misleading. The estimate of the gradient of the line can also be misleading in situations where outliers are present as the estimate is highly dependent on these few influential points. For example, on the random effects plot (Figure 31), the estimated gradient is 0.95, although it can be seen quite clearly that the gradient for the majority of the points is considerably less. Similarly the estimate of the slope of the fixed effect plot (Figure 30) is 2.06, due to the influence of the outliers.

In the more realistic situation where the within-study variances are different (Model (5)), although the fixed effect plot (Figure 32) and related tests still indicate quite clearly a deviation of the plot from normality, the message from the random effects plot (Figure 33) is far less clear. The plot shows only slight evidence of having a gradient less than one in the middle of the distribution, and in fact the estimated slope is approximately one. The p-values obtained from the Shapiro-Francia test and the Anderson-Darling test, although still detecting evidence against normality using

Figure 31: Random effects normal plot of q_i^* from a data set which is a mixture of two distributions with equal within-study variances (k=50) compared with the $N(0,1)$ line



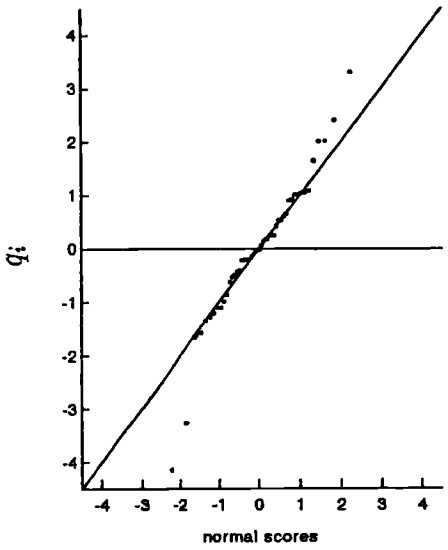
correlation=0.970

the random effects $q_{(i)}^*$ are far larger than when using the fixed effect $q_{(i)}$ (Table 26).

3.3.3 Investigation of the null distribution of the test statistics for a fixed effect plot

A potential problem exists with respect to the validity of the tests, since the theory behind the distributional result, that is that under the correct models $q_{(i)} \sim N\ 0,1)$ and $q_{(i)}^* \sim N(0,1)$, assume that θ is known. Hence, in practice where θ is estimated, the results are only approximate. In order to investigate the validity of the tests when parameters are estimated in the calculation of each $q_{(i)}$ and $q_{(i)}^*$, attention was focused on the null distributions of the test statistics. For the fixed effect $q_{(i)}$ firstly, under the null hypothesis the data follow a normally distributed fixed effect model, so that $\hat{\theta}_i \sim N(\theta, v_i)$ as $y_{il} \sim N(\theta, \sigma_i^2)$, where $v_i = \hat{\sigma}_i^2/n_i$ (Section 3.3.1). The FORTRAN

Figure 32: Fixed effect normal plot of q_i from a data set which is a mixture of two distributions with unequal within-study variances (k=50) compared with the $N(0,1)$ line

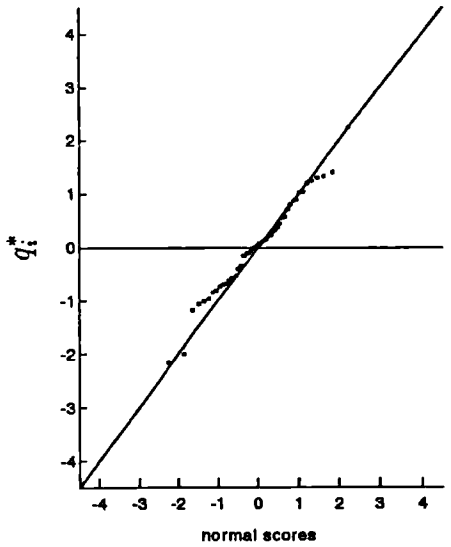


correlation=0.960

routine used here calculated four versions of the Anderson-Darling A^2 test statistic, together with the Shapiro-Francia W' test statistic. The number of times in 1000 normally distributed fixed effect model simulated data sets that each test statistic was observed to be more extreme than the 5% significance level was then counted, in order to obtain the Type-I error rate. If the null distribution of standard normality of the $q_{(i)}$ were to hold, then 5% ($\pm$ twice the standard error) of the 1000 tests would be significant. The number of trials in each data set was taken to be 20 in order to create a realistic meta-analysis situation.

There are then four different ways in which q_i can be calculated in the simulations, depending on whether θ or v_i or both, are estimated, and these are as follows:

Figure 33: Random effects normal plot of q_i^* from a data set which is a mixture of two distributions with unequal within-study variances (k=50) compared with the $N(0,1)$ line



correlation=0.976

(a) $q_i = (\hat{\theta}_i - \theta)/\sqrt{v_i}$

(b) $q_i = (\hat{\theta}_i - \theta)/\sqrt{\hat{v}_i}$

(c) $q_i = (\hat{\theta}_i - \hat{\theta}_f)/\sqrt{v_i}$

(d) $q_i = (\hat{\theta}_i - \hat{\theta}_f)/\sqrt{\hat{v}_i}$

where v_i now refers to a known within-study variance σ_i^2/n_i and $\hat{v}_i$ refers to an estimated within-study variance. Although it is only case (d) that is of practical concern, all four cases are considered in order to deduce whether the estimation of the parameters has any effect on the tests. Data from a correct model will strictly only produce a straight line through the origin with unit slope if it is assumed that both θ and v_i are known, that is in case (a). For all other cases this result is approximate and hence case (a) is the standard to which the others, and particularly (d), will be compared.

Furthermore, due to there being four different versions of the Anderson-Darling test (Section 3.2.2), the z_i used to calculate the A^2 statistic also take four different forms:

(1) $z_i = (q_i - 0)/1$
(2) $z_i = (q_i - 0)/\hat{\sigma}$
(3) $z_i = (q_i - \hat{\mu})/1$
(4) $z_i = (q_i - \hat{\mu})/\hat{\sigma}$

Hence, there are actually sixteen different versions of the Anderson-Darling test to be considered, together with four versions of the Shapiro-Francia test, that is one for each of (a)–(d). The version of the Anderson-Darling test of real interest for the situation under consideration is (1), since the interest lies in obtaining a test of standard normality and so the mean and variance of the distribution under the null hypothesis are both known.

Table 27: Results from the simulations under the null hypothesis that the data follow a normally distributed fixed effect model (that is when $q_{(i)} \sim N(0,1)$)

Version of q_i	Parameter		Type-I error (% significant from 1000 tests)				
	Overall effect (θ)	Within-study variance (v_i)	Version of A^2 test (1)	(2)	(3)	(4)	Shapiro-Francia *W'test*
(a)	known	known	5.0	5.8	5.2	5.6	5.4
(b)	known	estimated	4.8	4.9	4.3	4.5	4.3
(c)	estimated	known	**0.1**	**0.0**	4.9	6.4	6.6
(d)	estimated	estimated	**0.1**	**0.0**	4.8	4.7	4.7

It can be seen from the results that not all versions of the Anderson-Darling test produce the desired 5% error rate (Table 27). Focusing firstly on the practical

case where q_i is calculated using estimates of both θ and v_i, that is case (d), it can be seen that versions (1) and (2) of the Anderson-Darling test produce error rates which are far below 5% and are in fact approximately zero. This means that the test will be extremely conservative and, in practice, will be very low in power. The same results occur for case (c), but not for cases (a) and (b). The common factor linking cases (c) and (d) is that they are the two situations where θ is estimated in the calculation of $q_{(i)}$. Furthermore, the two versions of the test, (1) and (2), which produce low error rates are those where the mean of the distribution of the q_i under the null distribution is assumed to be known. Hence, the estimation of θ affects the performance of version (1) of the Anderson-Darling statistic and prevents it being of any practical use. The results for version (3) of the Anderson-Darling test are compatible with the 5% significance level in all cases.

The fact that the error rates for case (b) are around 5% for all versions of the Anderson-Darling test (Table 27) suggests that the estimation of the v_i does not cause any serious problems, at least for the example considered. Furthermore, version (3) of the Anderson-Darling test appears reasonable under case (d). Hence, version (3), that is where the null hypothesis tested is that of $H_0 : \theta \sim N(\mu, 1)$, may be used in practice. Although not being ideal, this version at least provides an improvement over the Shapiro-Francia test, in that the value of the variance of the $q_{(i)}$, or equivalently the gradient of the line on the fixed effect plot may be tested as well as the normality of the distribution. Hence, such a test is at least able to distinguish between a normally distributed fixed effect model and a normally distributed random effects model by detecting the increase in slope on the fixed effect plot with the latter.

Further investigation of version (1) of the Anderson-Darling test using q_i where θ and v_i are estimated, the case of practical interest, revealed that it is not the null hypothesis of standard normality of the q_i which is being tested . For an example where $v_i = v$ for all i, the null hypothesis actually being tested is $H_0 : \hat{\theta}_i \sim N(\theta, v)$,

where θ and v are unknown. By treating this as the null hypothesis, and therefore calculating the Anderson-Darling statistic using $z_i = (\hat{\theta}_i - \hat{\theta}_f)/\sqrt{\hat{v}}$ (i.e. using the q_i as the z_i), the correct 5% significance level is obtained. In a case where v_i=v for all i, this test of $\hat{\theta}_i \sim N(\theta, v)$ does check the distributional assumptions of a fixed effect model adequately. However, when the v_i are different each $\hat{\theta}_i$ has its own normal distribution $\hat{\theta}_i \sim N(\theta, v_i)$ and hence, the null hypothesis relating to the distribution of the $\hat{\theta}_i$ cannot be tested.

Finally, it is observed that the null distribution for case (c) for both the Shapiro-Francia test and version (4) of the Anderson-Darling test produce Type-I errors which are significantly larger than 5%. This may just be by chance, but it may also mean that the power of these tests is artificially increased. However, since case (c) is not of practical relevance and the increase is only slight, this issue was not investigated further. Also, the fact that the results for (d) are compatible with a value of 5% lends support to the view that they have occurred by chance.

3.3.4 Investigation of the null distribution of the test statistics for a random effects plot

Simulations similar to those described in Section 3.3.3 were used to consider the performance of the tests using the random effects $q^*_{(i)}$. This time, however, the data were generated under the normally distributed random effects model ((83) and (84) of Section 3.3.1) with σ^2_B=0.5. Only versions (3) and (4) of the Anderson-Darling test were considered here, since the previous investigations (Section 3.3.3) showed that versions (1) and (2) are obviously very conservative and lacking in power when θ is estimated, and are therefore of no practical use.

Confirmation of the validity of all of the tests when all parameters in q^*_i are known was obtained (Table 28). However, when the parameters θ, v_i, $i = 1, ..., k$ and

Table 28: Results from the simulations under the null hypothesis that the data follow a normally distributed random effects model (that is when $q^*_{(i)} \sim N(0,1)$)

Parameter		Type-I error (% significant from 1000 tests)		
Overall effect (θ)	Variance ($v_i + \sigma_B^2$)	Version of A^2 test (3)	(4)	Shapiro-Francia W' test
known	known	5.6	5.6	5.3
estimated	estimated	**0.7**	5.8	5.9

σ_B^2 are estimated in the calculation of q_i^*, even version (3) of the Anderson-Darling statistic is affected. Results for the null distribution indicate a significance level of under 1% rather than the required 5% (Table 28). This lowering of the type-I error rate must be due to the estimation of the between-study variance σ_B^2 in q_i^*, since it has been shown in Section 3.3.3 that the estimation of v_i alone does not apparently affect the test. Hence for the random effects components q_i^*, the only valid test is one of the null hypothesis $H_0 : q_i^* \sim N(\mu, \sigma^2)$ which is the same as that for the Shapiro-Francia W' test. The Shapiro-Francia test is not affected noticeably, if at all, by the estimation of the parameters in q_i^* (Table 28).

3.3.5 Conclusions from the simulations under the null hypothesis

The results obtained strictly only apply to normally distributed continuous outcome measures, as this is the type of data that were generated in the simulations. However, since θ is estimated in exactly the same way, that is by a weighted average of the individual estimates, for a binomial outcome, then it is likely that versions (1) and (2) of the Anderson-Darling test will again lose power. The simulations obviously indicate that the test of the null hypothesis $H_0 : q_i \sim N(0,1)$, that is where the null

distribution is completely specified, using the tables proposed by Stephens [106], is of little use in the practical situation where θ must be estimated in the calculation of q_i. An improvement over merely being able to test for normality is to test $H_0 : q_i \sim N(\mu, 1)$ and hence version (3) of the Anderson-Darling test is to be preferred when considering the distribution of the $q_{(i)}$. By testing this hypothesis, an advantage is obtained over the Shapiro-Francia test in that it enables a distinction to be made between a fixed effect model and a random effects model using the components $q_{(i)}$. However, on a random effects plot using $q^*_{(i)}$, version (4) of the Anderson-Darling test must be used, that is taking the null hypothesis to be $H_0 : q_i \sim N(\mu, \sigma^2)$, which is exactly equivalent to the null hypothesis of the Shapiro-Francia test.

In the simulations considered, since n_i was set to 50 for each study v_i will have been reasonably well estimated. Hence, further simulations for smaller values of n_i would be required to be able to make the generalisation that the estimation of v_i from individual study data does not noticeably affect the performance of the test.

3.3.6 Power of the tests for normality for fixed effect and random effects models

The power of the tests in relation to the fixed effect plots were investigated for models (1)−(3) (Section 3.3.2). Random effects plots were not considered since these models are all compatible with random effects models. For each example, 1000 repetitions were again simulated in order to investigate the performance of version (3) of the Anderson-Darling test, as well as the Shapiro-Francia test. Version (3) of the Anderson-Darling test was considered since it should be able to distinguish between a fixed effect model and a random effects model based on calculations using the fixed effect $q_{(i)}$. Thus, it should have a clear advantage over the Shapiro-Francia test for the three examples to be looked at. Each simulation was repeated twice, firstly as a standard for comparison against, taking the parameters θ and v_i to be known in the

calculation of q_i, and secondly, taking them to be estimated.

Table 29: Results of simulations looking at the power of the tests of normality using the fixed effect $q_{(i)}$ for data which follow three standard models (Section 3.3.2) when the overall treatment effect θ and the within-study variances v_i, $i = 1, ..., 20$, are known

Model	Power (% significant from 1000 tests)	
	Anderson-Darling A^2 test (Version(3))	Shapiro-Francia W' test
(1)	5.1	5.5
(2)	99.9	4.7
(3)	100.0	12.4

Table 30: Results of simulations looking at the power of the tests of normality using the fixed effect $q_{(i)}$ for data which follow the standard models (Section 3.3.2) when the overall treatment effect θ and the within-study variances v_i, $i = 1, ..., 20$, are estimated

Model	Power (% significant from 1000 tests)	
	Anderson-Darling A^2 test (Version (3))	Shapiro-Francia W' test
(1)	5.3	4.6
(2)	99.9	5.5
(3)	100.0	14.8

The results for the power of both tests under models (1), (2) and (3) are very similar whether θ and v_i, $i = 1, ..., k$ are known or whether they are estimated (Tables 29 and 30). Version (3) of the Anderson-Darling test exhibits good power when it comes to detecting a data set which follows a random effects model (Models

(2) and (3)) from the q_i of a fixed effect plot (Tables 29 and 30). This is because the variance of the distribution of each set of simulated q_i will be significantly greater than 1 and $H_0 : q_{(i)} \sim N(\mu, 1)$. The Shapiro-Francia test does find the random effects model with equal variances, Model (2), consistent with the null hypothesis of general normality. For Model (3), however, where the v_i are different, thus introducing some non-linearity into the plot, the power of the Shapiro-Francia test increases from the null level of 5%. However, the power is very low at only 12.4% for known parameters (Table 29) and 14.8% for estimated parameters (Table 30), since the test only detects deviations from linearity rather than the more obvious deviation of the variance from unity.

The Shapiro-Francia test cannot therefore effectively distinguish between the normally distributed fixed effect model and the normally distributed random effects model, based on the q_i from a fixed effect plot, whereas the Anderson-Darling test can. The plots may be used together and a random effects plot can consolidate the findings from a fixed effect plot. For example, a random effects plot which produces an approximate straight line with a gradient equal to 1, and corresponds to a fixed effect plot with a line with a gradient greater than 1, will support the conclusion that the data is well represented by a normally distributed random effects model.

3.3.7 Power of the tests for normality for data which conform to neither of the standard meta-analysis methods

As well as being able to distinguish between the two standard normally distributed meta-analysis models, it is useful to be able to detect data for which neither of these models is appropriate. However, there are many different alternative forms that such data could take, and hence only a brief exploratory investigation is practical here. A data set which consists of a mixture of observations obtained from both a fixed effect model and a random effects model is considered.

The mixed models (4) and (5) outlined in Section 3.3.2 were considered, but further variations were introduced whereby the number of random effects points in each set of data was varied. Hence for each model, (4) and (5), the number of points from a random effects model was taken to be 10, 20 and then 40. Again, both situations where θ and v_i, $i = 1, ..., k$, are known and where they are estimated were simulated. A further model (Model (6)) was also considered in order to investigate the effect of increasing the size of the between-study variance σ_B^2 relative to the within-study variances v_i. Model (6), therefore, uses the same value of v_i=v as model (4), but σ_B^2 is increased from 0.5 to 1.0.

Model (6) Data a mixture of fixed effect and random effects models

Large between-study variance σ_B^2=1.0 and v_i=v=0.1 for $i = 1, ..., k$

Number of points from a random effects model is 10, 20 and 40

Although neither the Shapiro-Francia test or version (3) of the Anderson-Darling test is affected by the estimation of θ and v_i under the null distributions, the power of both tests appears, in general, to increase slightly under the alternative models simulated here (comparing Tables 31 and 32). This may be due to the fact that additional variability is introduced into the q_i by the estimation of the parameters, thus meaning that there is an increase in the non-linearity of the normal plot. However, the interpretation of the tests remains unchanged, since the 5% error rate is maintained (Section 3.3.3). The Anderson-Darling test has consistently greater power than the Shapiro-Francia test when the parameters are known, except in the case of model 5) when the number of random effects points is 10. But even in this individual example, the powers are approximately the same (Table 31). In the realistic situation when the parameters are estimated, the power of both tests increases and so the Anderson-Darling test still remains more powerful than the Shapiro-Francia test (Table 32).

When there are only a few points from a random effects distribution, the power of the two tests tends to be similar with neither test performing particularly impressively (Tables 31 and 32). The power of the Anderson-Darling test does, however, increase as the number of random effects points increases, and, when there are 40 out of 50 points from a random effects model, the power is very high and approaches 100% for the values of σ_B^2 and v_i considered here. In contrast, the power of the Shapiro-Francia test generally decreases as the number of random effects points increases and is low when there are 40 random effects points. This contrast in the results observed may be easily explained by the fact that one test (W') is looking at only linearity $H_0 : q_{(i)} \sim N(\mu, \sigma^2)$ while the other (A^2) is concerned with the gradient of the slope as well $H_0 : q_{(i)} \sim N(\mu, 1)$. As the number of random effects points in the data increases, the variance of the distribution of the $q_{(i)}$, and equivalently, therefore, the gradient of the slope on the fixed effect normal plot, increases. Hence, version (3) of the Anderson-Darling test gains power as the variance gets increasingly larger than one. In contrast, particularly in the example where the within-study variances are equal (Models (4) and (6)), the plots will be more linear when there are 40 random effects points than when there are 10 such points. This is because 10 random effects points are more clearly seen as outliers among 40 fixed effect points than are 10 fixed effect points among 40 random effects points.

The power of both tests increases, in general, as the between-study variance of the random effects distribution increases from 0.5 to 1.0. This is to be expected since an increase in the between-study variance will lead to an increase in the variation observed in the data. Also, the power of both tests, although particularly that of W', tend to increase when the v_i are allowed to be different as opposed to being equal. The increase in power of the Shapiro-Francia test is due to the fact that the differing variances are an additional source of non-linearity on the fixed effect plot which may be detected by the test.

Table 31: Results of simulations looking at the power of the tests of normality using the fixed effect $q_{(i)}$ for data which follow a mixed model when the overall treatment effect θ and the within-study variances v_i, $i = 1, ..., 20$, are known

Model	Power (% significant from 1000 tests)					
	Number of random effects points					
	10		20		40	
	A^2	W'	A^2	W'	A^2	W'
(4)	29.9	22.9	77.1	20.2	99.6	7.8
(5)	34.8	34.9	80.8	41.0	99.8	22.2
(6)	66.2	49.4	98.5	46.0	100.0	8.7

A^2=Anderson-Darling test statistic

W'=Shapiro-Francia test statistic

When trying to detect in practice a set of data which is a mixture of the two standard models, it is not completely clear as to what approach to take. The Anderson-Darling test (version (3)) was shown to be a generally more powerful test than the Shapiro-Francia test for the purpose of detecting such data. Furthermore, a look at the plots may be helpful, although the random effects plots may sometimes be misleading in that it may suggest that the random effects model is adequate when in fact the test shows it not to be (Section 3.3.2).

3.4 Practical Examples

Data sets will now be considered to provide examples where the techniques described in Sections 3.1 and 3.2 are used as an aid to the interpretation and investigation of heterogeneity. Section 3.4.1 looks at the reduction in blood pressure in the mild hypertension trial (Section 1.3.2), an example where the normal plots produce clear

Table 32: Results of simulations looking at the power of the tests of normality using the fixed effect $q_{(i)}$ for data which follow a mixed model when the overall treatment effect θ and the within-study variances v_i, $i = 1, ..., 20$, are estimated

Model	Power (% significant from 1000 tests)					
	Number of random effects points					
	10		20		40	
	A^2	W'	A^2	W'	A^2	W'
(4)	38.4	27.1	80.3	24.1	99.8	9.3
(5)	43.2	37.1	79.6	41.0	99.8	23.1
(6)	68.2	52.3	98.4	46.7	100.0	13.2

A^2=Anderson-Darling test statistic

W'=Shapiro-Francia test statistic

pictures owing to the large number of observations available. Section 3.4.2 considers an example, the diuretics trials meta-analysis (Section 1.3.1), where the plots are of less use for checking distributional assumptions because of the small number of trials. However, the example does show how the plots may be useful for investigating sources of heterogeneity, and the Galbraith plot [66] is also presented as a further way of displaying meta-analysis data.

3.4.1 Mild Hypertension Trial

The assumption of homogeneity of treatment effect across all centres in the mild hypertension trial was tested formally using the Q statistic (Section 1.6). The test was carried out for the reduction in both systolic blood pressure (SBP) and diastolic blood pressure (DBP) on the results for the treatment and placebo groups separately, and also for the difference in blood pressure reduction between these two groups.

Centre 1 was excluded in these analyses for reasons given in Section 1.3.2.

Table 33: Results of the test for heterogeneity for diastolic blood pressure reduction between entry to the MRC mild hypertension trial and a year after entry

Outcome	Statistic for heterogeneity (Q)	Degrees of freedom (df)	Q/df	p-value
Placebo group	1131.51	188	6.0187	<0.0001
Treatment group	953.79	188	5.0734	<0.0001
Difference between placebo and treatment	278.12	188	1.4794	<0.0001

Table 34: Results of the test for heterogeneity for systolic blood pressure reduction between entry to the MRC mild hypertension trial and a year after entry

Outcome	Statistic for heterogeneity (Q)	Degrees of freedom (df)	Q/df	p-value
Placebo group	1052.06	188	5.5961	<0.0001
Treatment group	760.78	188	4.0467	<0.0001
Difference between placebo and treatment	246.28	189	1.3031	<0.005

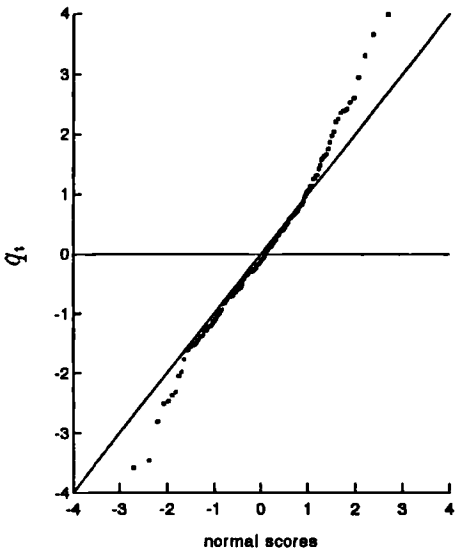
The results show strong evidence of a lack of homogeneity across centres in all cases (Tables 33 and 34). However, there is far greater heterogeneity within each group individually than there is when the two groups are compared by taking the difference in means. The extent of the heterogeneity, summarised by Q/degrees of freedom, in the results for DBP (Table 33) is somewhat greater than the evidence of heterogeneity in those for SBP (Table 34).

The results for the q-q plots will now be discussed for DBP only, as those for SBP provide an almost identical picture. When looking at the difference in reduction

of blood pressure between treatment and control groups using a fixed effect q-q plot, the majority of the data falls along the line of identity and it is only in the tails of the distribution where there is any deviation from this line (Figure 34). This plot is similar to those obtained from simulations of data from a mixture of two distributions (Figures 30 and 32). The Shapiro-Francia W' test ($p \simeq 0.01$) and the Anderson-Darling test ($p<0.01$) indicate that there is strong evidence against normality. The estimated slope of the regression line of 1.47 is not very informative in this example as it can clearly be seen that the plot is made up of two different groups of points and the estimate is influenced by the points in the tails of the distribution of $q_{(i)}$. When the corresponding random effects plot is looked at, there is no great visual evidence of a lack of fit of the random effects model, although it should be noted that the middle section of the plot does have a gradient less than one (Figure 35). This plot may suggest that the random effects model with the normality assumptions is a reasonable representation of the data. The estimate of the slope is only slightly less than one (0.99) and therefore strengthens the view that the random effects model is a reasonable fit to the data. However, as was seen with the simulated examples, the estimate of the slope may be highly dependent on extreme and outlying values. Rather surprisingly, the Shapiro-Francia test provides stronger evidence of non-normality from the $q^*_{(i)}$ than it did from the $q_{(i)}$ and produces a p-value of 0.007. The explanation may be that there are greater deviations from linearity in the centre of the distribution which the test is sensitive enough to detect, and, furthermore, the change of scale on the random effects plot may be rather deceptive. Hence the random effects plot does not show the model violation as clearly as the fixed effect plot, but according to the test based on the $q_{(i)}$ it provides stronger evidence against normality than the test based on $q^*_{(i)}$.

The fixed effect plot for each individual group (treatment and placebo), in contrast, both produce straight lines through the origin with gradients steeper than one (Figure 36). The estimate of this slope for the treatment group is 5.04 and

Figure 34: Fixed effect normal plot of q_i for the difference in the reduction in diastolic blood pressure between the treatment and control group in the mild hypertension trial compared with the $N(0,1)$ line

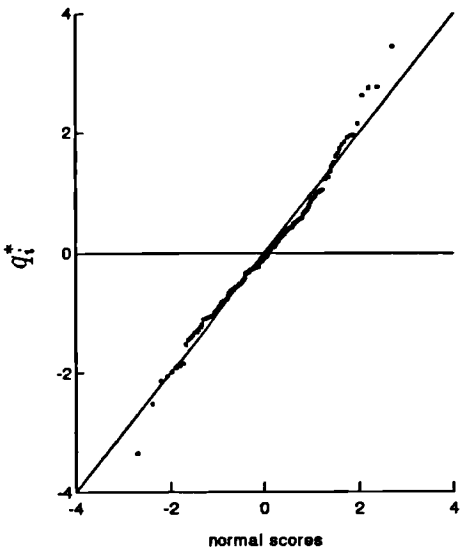


correlation=0.990

this evidence suggests that there is great heterogeneity present which is distributed throughout the centres. The results of the Shapiro-Francia test ($p \simeq 0.008$) and version (3) of the Anderson-Darling test ($p<0.01$) suggest that the plot is not in fact linear and that there is significant evidence against normality. These results, together with the shape of the plot, indicate that the sample may be from a random effects model with different within-study variances. The random effects plot (Figure 37) reinforces this view as the points fall along the line representing the N(0,1) distribution. The W' test ($p \simeq 0.08$) and the Anderson-Darling test ($p<0.1$) provide some evidence of non-normality, but this is not overwhelming. Also, the estimate of the gradient is very close to unity and hence the random effects model may be an acceptable approximation to the data obtained from the treatment group.

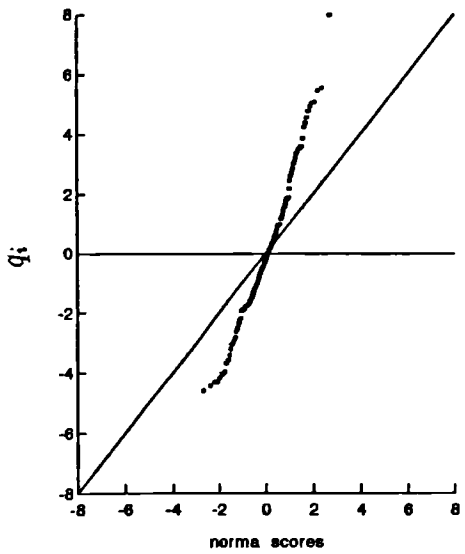
In the case of the difference in blood pressure reduction between the two groups, it is possible to remove the heterogeneity from the data by omitting a small

Figure 35: Random effect normal plot of q_i^* for the difference in the reduction in diastolic blood pressure between the treatment and control group in the mild hypertension trial compared with the $N(0,1)$ line



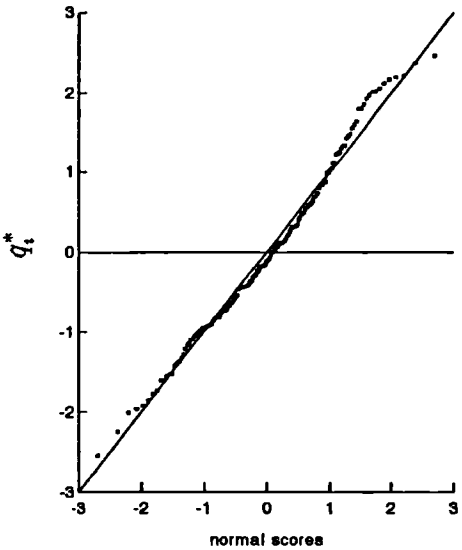
correlation=0.989

Figure 36: Fixed effect normal plot of q_i for the reduction in diastolic blood pressure in the treatment group in the mild hypertension trial compared with the $N(0,1)$ line



correlation=0.990

Figure 37: Random effects normal plot of q_i^* for the reduction in diastolic blood pressure in the treatment group in the mild hypertension trial compared with the $N(0,1)$ line



correlation=0.994

number of the outlying centres, that is those in the tails of the q-q plot. In contrast, for the individual groups (treatment and placebo), the heterogeneity is distributed throughout all the centres. In order to deduce the centres which are the source of the heterogeneity for the difference between groups in the reduction of blood pressure, 'outlying' centres were excluded one at a time, in order of decreasing size of q_i^2. As each centre was removed Q was recalculated, using a new estimate of θ each time, and this procedure continued until enough centres had been removed to produce a value of Q which corresponded to a p-value of 0.1 or above. The value of the degrees of freedom of the χ^2 distribution to which Q is compared is obviously reduced by one each time a centre is removed.

Table 35: Centres are removed in turn, starting with the most heterogeneous, until $p > 0.1$ for the test for heterogeneity Q for the difference in the reduction in diastolic blood pressure between the treatment and placebo groups

Centre removed	q_i^2 for centre removed	Statistic for heterogeneity (Q)	Degrees of freedom ($k-1$)	$\chi^2_{(k-1)}$ (p=0.1)
		278.12	188	214
118	15.94	262.18	187	213
65	13.35	248.83	186	212
162	12.82	236.02	185	211
186	11.97	224.05	184	210
36	10.96	213.09	183	209
176	8.73	204.36	182	208

k=189=number of centres in study

Heterogeneity, for the difference in the reduction of blood pressure between groups, can be removed by omitting only 6 centres for each of systolic and diastolic blood pressure (Tables 35 and 36). It is not the same set of centres that contribute to the heterogeneity for both systolic and diastolic blood pressure outcomes, although there is some overlap. The reasons for the large contributions being made to Q by these particular centres were identified and are recorded in Tables 37 and 38. The possibility of investigating clinical reasons behind such results in these centres then exists, but is not pursued here.

Plots of $q_{(i)}$ for DBP against $q_{(i)}$ for SBP show a positive correlation (Figure 38). Hence, larger than average reductions in DBP tend to be accompanied, as would be expected, by correspondingly larger than average reductions in SBP. The plots of $q_{(i)}$ for the placebo group against the treatment group for both DBP (Figure 39) and SBP clearly show how there is great variation across centres within each treatment

Table 36: Centres are removed in turn, starting from the most heterogeneous, until $p > 0.1$ for the test for heterogeneity Q for the difference in the reduction in systolic blood pressure between the treatment and placebo groups

Centre removed	q_i^2 for centre removed	Statistic for heterogeneity (Q)	degrees of freedom ($k-1$)	$\chi^2_{(k-1)}$ (p=0.1)
		246.28	188	213
3	7.84	238.43	187	212
18	7.45	230.99	186	211
48	7.10	223.88	185	210
118	6.92	216.96	184	209
186	6.47	210.49	183	208
30	6.46	204.03	182	207

k=189=number of centres in study

group, that is in the x and y direction on the plot. They also indicate that a strong relationship exists between the blood pressure reduction in the placebo group and that in the treatment group. Hence, centres with larger than average reductions in the treatment group also tend to have large reductions in the placebo group.

Hence, this information indicates that there is an important 'centre effect' in this set of data. The effect is seen in the individual groups, but when the difference is taken this 'centre effect' is largely cancelled out, leaving the heterogeneity confined to only a handful of centres. It is, furthermore, interesting to note the presence of a 'placebo effect' in this study, whereby a mean reduction in blood pressure occurs in the group of patients who received only a placebo rather than an active drug. There is a large variation in the 'placebo effect' across centres, with observed average reductions in blood pressure in some centres being extremely large. The 'placebo effect' may result from the psychological effect on the patients of participation in a

Table 37: Reasons for the large contributions to heterogeneity of the centres removed for the difference in the reduction of diastolic blood pressure (DBP) between the treatment and placebo groups

Centre	No. patients in centre	Difference	Variance of difference	Mean reduction (placebo)	Mean reduction (treatment)
118	61	14.47	5.26	3.94	18.41
65	146	10.40	1.92	-1.00	9.40
162	29	-7.47	12.79	8.89	1.41
186	142	0.34	2.06	7.75	7.86
36	79	12.07	4.15	-1.60	10.47
176	35	15.50	11.88	-0.16	15.34

Difference=difference in mean DBP reduction between treatment and placebo groups

(Overall average difference in mean DBP reduction between groups=5.1mmHg)

Mean reduction=mean reduction in DBP in single group

Centre	Reason for large q_i^2
118	Large treatment effect
65	Placebo group has negative difference and small variance
162	Placebo effect much larger than treatment effect
186	No difference in effect between groups and small variance
36	Placebo group has negative difference
176	Large difference between groups

Table 38: Reasons for the large contributions to heterogeneity of the centres removed for the difference in the reduction of systolic blood pressure (SBP) between the treatment and placebo groups

Centre	No. patients in centre	Difference	Variance of difference	Mean reduction (placebo)	Mean reduction (treatment)
3	104	20.77	13.00	1.16	21.39
18	19	-8.27	48.23	13.67	5.40
48	107	1.96	10.72	25.00	26.96
118	61	21.22	16.02	12.41	33.63
186	142	3.27	8.51	22.32	25.59
30	105	1.95	11.84	20.11	22.05

Difference=difference in mean SBP reduction between treatment and placebo groups (Overall average difference in mean SBP between groups=10.4mmHg)
Mean reduction=mean reduction in SBP in single group

Centre	Reason for large q_i^2
3	Small placebo effect and quite large treatment effect
18	Placebo effect much larger than treatment effect
48	Small difference in effect between groups
118	Large treatment effect
186	Small difference in effect between groups
30	Small difference in effect between groups

Figure 38: Plot of q_i for diastolic blood pressure against systolic blood pressure for the difference between the treatment and the placebo group (correlation=0.539)

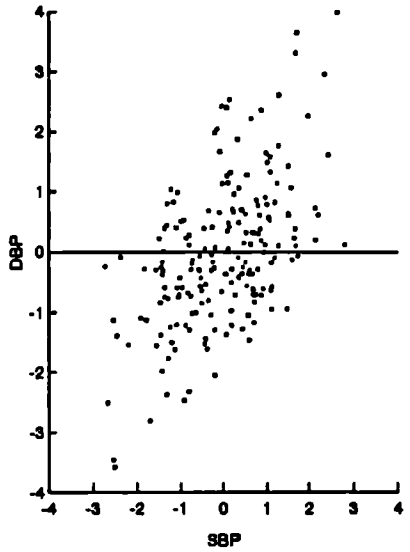
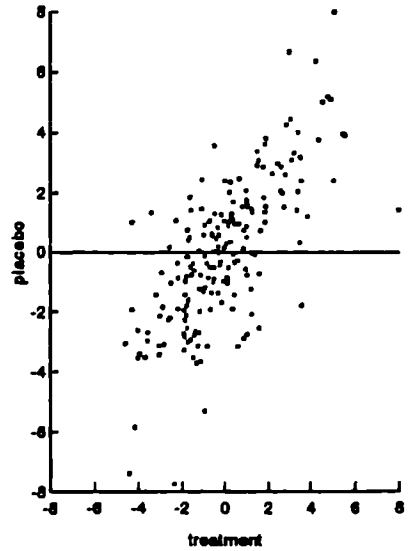


Figure 39: Plot of q_i for placebo group against treatment group in the mild hypertension trial (correlation=0.712)



trial and their being put on a course of tablets, as well as 'regression to the mean' and some patients being changed to an active treatment. The heterogeneity in the treatment and the placebo groups could be partially explained by, for example, a 'nurse effect', where some centres may have had a particularly reassuring research nurse. Additionally, the standard of further care may have varied between centres and this could have lead to heterogeneity, or the characteristics of the patients may have varied between centres too, with certain groups possibly being more responsive to treatment than others.

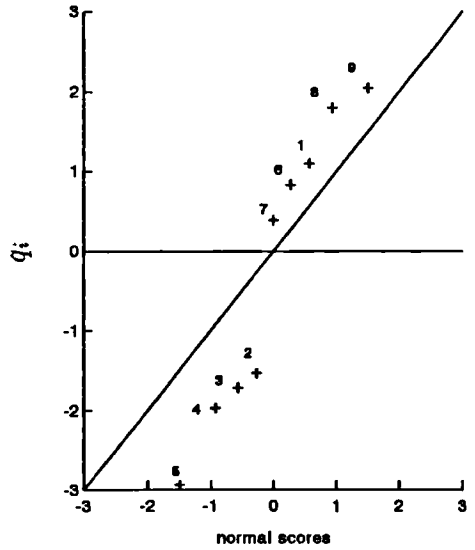
In this example, the plots have provided information regarding the possible distribution of the data and hence, the suitability of the standard meta-analysis models. Information was also gained about the location and distribution of the heterogeneity and possible outlying centres could be identified for further investigation.

3.4.2 Diuretics Trials Meta-Analysis

In contrast to the multicentre trial of the previous section, where there were 189 observations on the q-q plot, actual meta-analyses tend to contain only a limited number of trials. In such situations the q-q plots are not so informative with regards to the distribution of the data. The diuretics trials meta-analysis (Section 1.3.1) illustrates the problem that is likely to be encountered since it produces q-q plots with only nine points. The fixed effect plot (Figure 40) clearly indicates the presence of heterogeneity within the data and that the $q_{(i)}$ are not standard normal as the gradient of the plot is steeper than one, with the actual estimate of the regression line being 3.55. Furthermore, the plot suggests the possibility of there being two separate groups of trials. The first group of trials, which have individual treatment effect estimates greater than the overall mean (trials 1, 6, 7, 8, 9), fall closer to the line of identity than the group of trials with estimates smaller than the overall mean (trials 2, 3, 4, 5). The information thus gained may then be helpful when investigating the

reasons underlying the heterogeneity. For example, trial 5, which makes the largest contribution to the statistic for heterogeneity can be identified as a possible outlier because it produces one of the largest treatment effects. Investigation reveals that this is the only trial which uses the drug Bendroflumethiazide for the treated group and, furthermore, that the entry criterion is 30 weeks or more into the pregnancy, which is later than in nearly all the other trials. Hence, it may be that the result observed in this trial was due to its having these different characteristics in design.

Figure 40: Fixed effect normal plot of q_i for the diuretics trials meta-analysis compared with the $N(0,1)$ line

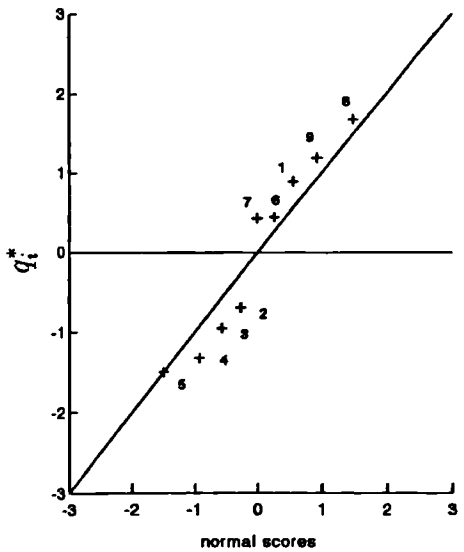


If an explanation for the two groups of trials observed were found, then it might be more reasonable to carry out a separate meta-analysis on each subgroup or to use regression modelling with an indicator for each subgroup to account for the variation. However, no characteristic about which information could be gained from the published paper, such as type of drug, drug regimen, type of patient, entry criteria, was found to explain the apparent bisection of the trials. Hence, it may be that the reasons for the extra variation are too complicated to sort out. Alternatively,

further investigation, using information from the original trials, may be required.

The evidence from the random effects q-q plot (Figure 41) may suggest that the random effects model is a more reasonable fit to these data as all points have been pulled in towards the line of identity. It is particularly noticeable that the point representing trial number 5 has been pulled in very considerably and now falls on the line of identity. This is due to its having a small within-study variance compared to the between-study variance and so being shrunk by a large proportionate amount. On the other hand, the $q_{(i)}$ for trial 7, for example, stays approximately the same on both plots, because it has a relatively large within-study variance and has an individual estimate close to the estimate of overall treatment effect. However, because there are so few points, it is still not clear whether the points do in fact form a line through the origin or whether there are still two separate lines indicating that there are still two separate groups of trials. The estimate of the gradient is, however, still greater than unity at 1.45.

Figure 41: Random effects normal plot of q_i^* for the diuretics trials meta-analysis compared with the $N(0,1)$ line

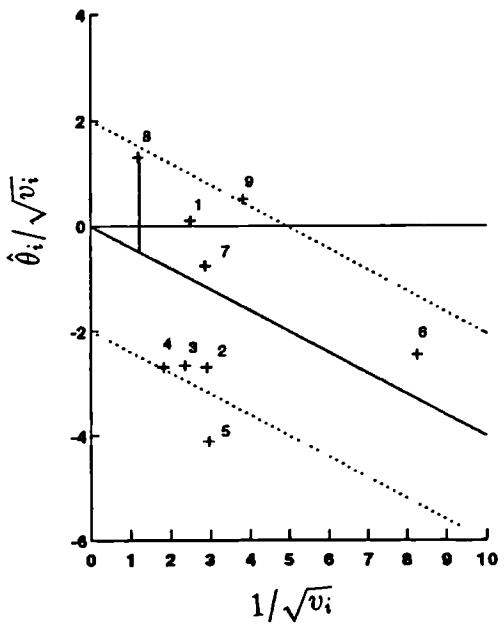


correlation=0.972

The tests for normality, in this example with so few points, are less helpful, since for neither $q_{(i)}$ or $q^*_{(i)}$ do either provide any evidence against normality. This is, however, likely to be due to the lack of power of the test with only nine points.

With respect to identifying sources of heterogeneity and displaying them, the q-q plot provides the same information as the 'radial plot' described by Galbraith [66]. The Galbraith plot is a scatter plot of $y = \hat{\theta}_i/\sqrt{v_i}$ against $x - 1/\sqrt{v_i}$ where a single point is plotted for each trial (Figure 42). A radial log odds ratio scale can then be used to read off values represented by lines through the origin and the point of interest, (x_i, y_i). The horizontal axis of the graph, therefore, corresponds to an odds ratio of 1. Such a plot can be simplified and drawn without the radial scale, although still providing the same useful information.

Figure 42: Galbraith plot of the diuretics trials meta-analysis



Trials having estimates of treatment effect with small standard errors, which are therefore those which provide the most information, lie well away from the origin due to their large x-coordinates. On the other hand, less informative studies produc-

ing estimates with large standard errors cluster near to the origin. Hence, the points falling away from the origin look naturally most informative [66], which indeed they are. It can therefore easily be observed from where the largest amount of information is derived.

Since the gradient x/y is equal to $\hat{\theta}_i/\sqrt{v_i}/1/\sqrt{v_i} = \hat{\theta}_i$, a line going through the origin with gradient $\hat{\theta}_f$ represents the fixed effect estimate of the overall log odds ratio of the meta-analysis. If the studies are homogeneous, then the points will scatter homoscedastically, with unit standard deviation, about this line. Furthermore, lines representing two standardised units either side of the overall odds ratio line may also be drawn on the diagram to aid interpretation. The further away the point is from the line representing the overall odds ratio, the more heterogeneous it is. The vertical distance from the point representing trial i to the line representing $\hat{\theta}_f$ is equal to q_i (Figure 42), the value plotted on the fixed effect normal plot for study i. The point, for trial i, on the Galbraith plot is $\hat{\theta}_i/\sqrt{v_i}$ and the corresponding point on the odds ratio line is $\hat{\theta}_f/\sqrt{v_i}$ and so the distance between them is $(\hat{\theta}_i - \hat{\theta}_f)/\sqrt{v_i} = q_i$.

It can be seen from the Galbraith plot for the diuretics trials data (Figure 42) that there is substantial heterogeneity present, since most of the points lie well away from the line representing $\hat{\theta}_f$. As with the normal plot, trial 5 is seen to be an outlying observation with a highly negative estimate of the log odds ratio, together with a relatively small variance, meaning that the point falls well away from the line representing $\hat{\theta}_f$.

In the discussion of DeMets [7], Peto suggested an alternative but similar plot to that of Galbraith, based on the Peto method of meta-analysis, namely a scatter plot of $(O - E)$ against V (notation in Section 1.5.3), in which small trials cluster near to the origin and the informative trials are far to the right of the plot. Galbraith [66] points out that a close approximation to the 'radial' plot is actually a scatter plot of $(O - E)/\sqrt{V}$ against $\sqrt{V}$.

Both the Galbraith plots and the q-q plots can help to identify outliers and subgroups of homogeneous trials, which may then lead to further investigations. However, it should always be remembered that any explanations derived from such observations are post-hoc and should be interpreted cautiously. Furthermore, there may sometimes be more than one feasible explanation or none. Hence, a sensible way of using the findings from such investigations may be to generate hypotheses for future trials.

3.5 Conclusion

There is clearly no single superior method for checking the distributional assumptions of meta-analysis models. A combination of q-q plots and both the Shapiro-Francia and versions (3) and (4) of the Anderson-Darling tests for normality should usually be adequate to check the distributional assumptions when a reasonably large number of trials is available. The plots are perhaps more useful in the case of multicentre trials when more data points tend to be available (Section 3.4.1), so the shape of the plot is clearer. In many meta-analyses there will not be enough studies to make full use of these techniques. However, the plots can still provide useful information for identifying outlying studies and groupings within the studies and therefore aiding in the interpretation of heterogeneity (Section 3.4.2).

It must be remembered that the version of the Anderson-Darling statistic to be used in practical circumstances for the $q_{(i)}$ relating to a fixed effects plot is that assuming a null hypothesis of normality with unknown mean and variance equal to 1. For the random effects plot no advantage over the Shapiro-Francia test is gained by using the Anderson-Darling statistic as both only test the null hypothesis of general normality. If it is assumed that the null distribution is completely specified as standard normal, the test is lacking in power, and, if used, the power to detect non-normality will be greatly reduced.

Nevertheless, the random effects plots do provide a check of the normality of the random effects as well as the data. Hence, if it is the distribution of the random effects, that is $\theta_i \sim N(\theta, \sigma_B^2)$, which is of primary concern, then a random effects plot will be useful. However, it has been noted [108] that the unweighted random effects normal plot used here may be inefficient for such purposes. Dempster and Ryan [108] suggest the use of weighted normal plots as an improvement to the straightforward unweighted plots (Section 3.1.2) to check the normality assumptions of the random effects in linear models. The method can easily be adapted to the case of a random effects meta-analysis. It has been shown [108] that weighted plots are more sensitive than unweighted plots to certain departures from the assumed distribution. For example, it is more sensitive for detecting a misspecified variance, for detecting outliers among the random effects and for detecting when the distribution has a long tail. However, such plots do have greater pointwise variability than the unweighted plots [108].

The standard unweighted plot gives equal weight to each trial, even though some study estimates will contain more information about the random effects than others [108]. Hence, the weighted method involves the assignment of greater weight to those observations for which σ_B^2 accounts for a larger portion of the overall variance $(v_i + \sigma_B^2)$. A simple choice of weights is $w_i^* = 1/(v_i + \sigma_B^2)$ and then

$$F_k^*(q) = \sum_{i=1}^{k} I(q - q_i^*) w_i^* / \sum_{i=1}^{k} w_i^* \qquad (86)$$

In order to make these plots equivalent to the unweighted plots, where Blom's modification [102] is used, adjustments to $F_k^*(q)$ must be made [108]. As with the previous unweighted plots, under the correct model the q-q plot should be a straight line through the origin with unit gradient.

Exploratory investigations were carried out regarding the weighted normal

plot using simulated data of the kind for which the weighted plot should be more sensitive, for example, where there are outliers among the random effects. However, no obvious advantages were seen and furthermore the weighted plots were found to be very similar to the unweighted random effects plots. Also, given that the weighted plots involve more complex calculations to produce, little advantage was seen in the meta-analysis situation.

If investigations into the distributional assumptions show that either a fixed effect or a normally distributed random effects model is reasonable, then the confidence in the results obtained from the standard methods will be increased, However, a problem which requires further research is that of what to do if neither model is found to be satisfactory. Investigation is required into the robustness of the standard results to deviations from the assumed models, and also into alternative ways of modelling such data. One possibility is to use a non-parametric distribution for the random effects component such as in the method proposed by van Houwelingen et al. [45].

4 Power of the Test for Heterogeneity in Meta-Analysis

The test for heterogeneity of treatment effects across studies in a meta-analysis is often said to have 'low power' [76], but the actual power is rarely quantified. However, a simulation study was carried out, considering heterogeneity in k 2×2 tables, which showed the low power of heterogeneity tests in general, particularly when data are sparse [109]. Hence, this chapter investigates the power of the test for heterogeneity in meta-analyses. The assessment is based on the usual test statistic for heterogeneity, using Q (Section 1.6) which is referenced to a χ^2_{k-1} distribution.

The statistic Q is based on the assumption that each weight w_i is known rather than estimated and is equal to the reciprocal of the variance v_i of the individual trial estimate. However, in a practical situation, estimated weights must be used, which are usually derived from the estimated variances. The work in this chapter is carried out under the assumption that the w_i are known, but the issue of estimating w_i will be pursued in Chapter 5.

Section 4.1 describes the methods and the strategy used for this investigation of the power of the test for heterogeneity and Section 4.2 presents and discusses the results. It is illustrated how the power of the test varies with the between-study variance σ^2_B, the number of trials k in the meta-analysis and the weight allocation. An alternative statistic to Q for testing heterogeneity is then considered in Section 4.3 and conclusions are drawn in Section 4.4.

4.1 Methods

The power of the test for heterogeneity is dependent on the distribution of the values of Q. The expectation of the test statistic Q may easily be obtained for any given random effects meta-analysis model, providing it is assumed that the weights are known. The expectation of Q [38] is given by

$$E(Q) = (k-1) + \sigma_B^2 \left(\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i} \right) \tag{87}$$

Analytic results for the power of the test are, however, not simple to produce. In general, however, the larger the expected value of Q, for a given number of degrees of freedom, the greater the power. However, differences in the distribution of Q will mean that if two examples have the same expected value of Q, the power may still be different. Therefore, the power of the test was investigated using computer simulation methods; these were again written in FORTRAN and were based on the models defined by (82) and (83) of Section 3.3.1. Hence, the results obtained in this chapter are all based on assuming the parameters are known rather than estimated. This chapter, therefore, concentrates on quantifying the theoretical power of the test, while Chapter 5 extends this work to look at the effect that estimating the weights has on the power of the test for heterogeneity.

In this chapter it is always assumed, for simplicity, that the heterogeneity to be detected takes the form of a normally distributed random effects model. Hence, all the results obtained relate to the power of the test in detecting such heterogeneity. An extension of the work would be to look at the power of the test for detecting heterogeneity which takes alternative forms, for example, heterogeneity caused by a few outlying points.

There are then three characteristics of a meta-analysis data set which will

have an effect on the expectation of Q and therefore on the power of the test for heterogeneity. They are (a) the extent of heterogeneity present, that is the value of the between-study variance σ_B^2, (b) the number of studies included in the meta-analysis k, and (c) the weights w_i allocated to the individual studies. Factor (c) is the most complicated to investigate, since there are endless different combinations of weights that could be considered. However, by making certain simplifications, the behaviour of the test statistic, with respect to weight allocation, could be investigated reasonably fully.

The strategy employed to investigate the three factors (a)–(c) is now described. For each meta-analysis example considered, simulations were carried out for a variety of values of the between-study variance and 1000 repetitions were performed for each value of σ_B^2 (σ_B^2=0, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5). Hence, both the mean value of the simulated Q and the power of the test could be plotted against the between-study variance σ_B^2. Under the null distribution 5% (± twice the standard error) of the 1000 tests should produce a statistic larger than $\chi^2_{k-1,0.95}$. Initially, the weights allocated to each trial within each meta-analysis were kept equal, w_i=w for all i, in order that the effect of the number of trials k and the total amount of information $\sum_{i=1}^{k} w_i$ be investigated. Simulations were carried out with w=10, and the effect of varying the number of trials (k=5, 10 and 20 were used), and hence also the total weight, was observed. Further simulations to investigate the effect of changing k (k=5, 10 and 20) were carried out where the total information was kept fixed, $\sum_{i=1}^{k} w_i$=100 here. Hence, changing k implies that w changes. Similarly, to investigate the effect of changing the amount of total information for a given k, the number of studies was fixed at k=10 and the total weight was varied ($\sum_{i=1}^{k} w_i$=50, 100 and 200 were used), implying that w varied too.

The behaviour of the test statistic, in relation to the changing allocation of weight, was then considered, without any loss of generality, by fixing the value of

$\sum_{i=1}^{k} w_i$, while letting the individual w_i vary. This is illustrated by noting from equation (87) that $E(Q)$ plotted against σ_B^2 produces a straight line with intercept $(k-1)$ and gradient $W = (\sum_{i=1}^{k} w_i - (\sum_{i=1}^{k} w_i^2 / \sum_{i=1}^{k} w_i))$ and assuming that each w_i is fixed, then if $\sum_{i=1}^{k} w_i$ is multiplied by any factor x, the sums $x\sum_{i=1}^{k} w_i = \sum_{i=1}^{k} xw_i$ and $\sum_{i=1}^{k}(xw_i)^2 = x^2 \sum_{i=1}^{k} w_i^2$ are obtained. This means that when the expectation of Q is calculated, the gradient of the plot of $E(Q)$ against σ_B^2 is simply multiplied by x,

$$E(Q) = (k-1) + \sigma_B^2 x \left(\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i} \right) \tag{88}$$

Hence, the results for any sum of weights $\sum_{i=1}^{k} xw_i$ differ only in the scale of the plot of $E(Q)$, although how this relates to the results for the power of the test is not clear.

For the purpose of this investigation, k=10 was used and the total information $\sum_{i=1}^{k} w_i$ was fixed at 100. A further simplification was made, that of restricting the investigation to consider examples where all trial weights, apart from w_1, are equal to each other. Three cases were chosen to represent a large range of possible values of w_1. These cases were: $w_1 = 10$ (i.e. the situation with all w_i equal), $w_1 = 50$ and $w_1 = 90$ (i.e. an extreme situation with a single study dominating the meta-analysis). The fact that, for the equal weighting case, w_i=10 for $i = 1, ..., k$ means that v_i=0.1 for $i = 1, ..., k$. Hence, the range of values of σ_B^2 investigated covers cases where the between-study variation is smaller, equal to and larger than the variation within each individual study.

The mean values of Q, from 1000 repetitions, were obtained from the simulations together with the power of the test for heterogeneity, that is the number of times in the 1000 repetitions where a test statistic was produced which was significant at the 5% level. The true expected value of Q was also calculated analytically for every example using formula (87), and this value was compared to the simulated mean of

Q in order to provide a check on the validity of the simulations.

4.2 Results

The results relating to the three factors identified as affecting the power of the test, outlined in Section 4.1, are each described in Sections 4.2.1–4.2.3. Alternative ways of viewing the results for the power of the test, in order that the practical implications be highlighted, are then considered in Section 4.2.4.

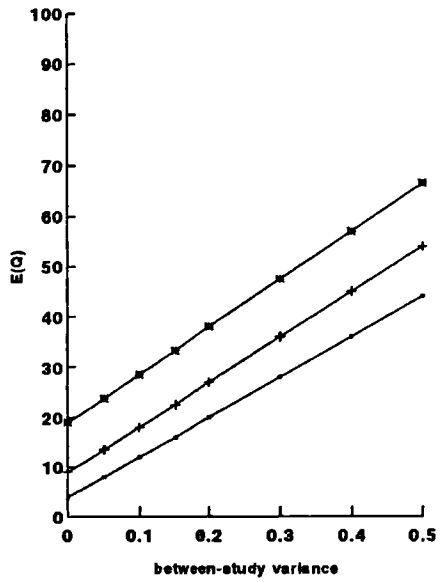
4.2.1 Power and the between-study variance

The value of $E(Q)$, for a given k, increases as the extent of the heterogeneity increases, that is as the between-study variance gets larger (Table 39). As has already been noted, $E(Q)$, when plotted against σ_B^2, will form a straight line with intercept $(k-1)$ and gradient $W = (\sum_{i=1}^{k} w_i - \sum_{i=1}^{k} w_i^2 / \sum_{i=1}^{k} w_i)$ (Figure 43). Figure 43 provides examples of the plots where the number of trials k varies, but where $\sum_{i=1}^{k} w_i$ is fixed and within each example the weights are equal.

The 95% confidence intervals, calculated for the simulated mean Q, may be used to check the validity of the simulations. All such intervals contained the true expected value of Q (Table 39), thus indicating that the simulated results obtained for the power of the test should be reliable.

For the behaviour of the power of the test, it can be seen that, following on from the pattern obtained from E(Q), the power also increases with increasing heterogeneity. The plots of power against the between-study variance take the familiar form of a power curve, starting from 5% when there is no heterogeneity and then levelling out as 100% power is reached (Figure 44).

Figure 43: Expectation of the Q statistic against the between-study variance for different numbers of trials k when $\sum_{i=1}^{k} w_i$=100 and the weights are all equal



Key

. k=5, W=80

\+ k=10, W=90

* k=20, W=95

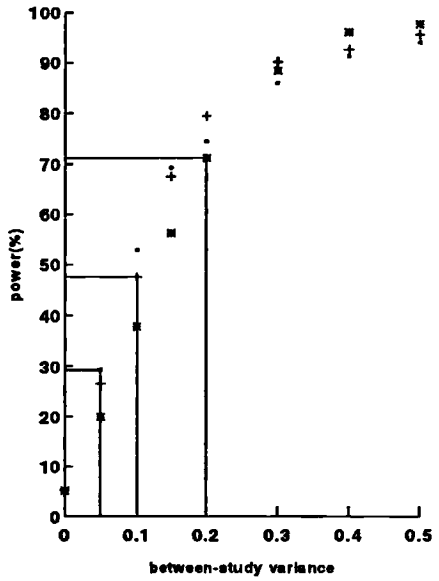
1000 simulations at each point

Table 39: Mean observed values of the test statistic for heterogeneity Q from the simulations compared to the true expected values where the total sum of weight is equal to 100

σ_B^2	Weight allocated to trial 1 w_1 ($w_2 = ... = w_{10}$)					
	10		50		90	
	$w_2, .., w_{10}$=10, W=90		$w_2, .., w_{10}$=5.6, W=72.2		$w_2, .., w_{10}$=1.1, W=18.9	
	Analytic	Observed	Analytic	Observed	Analytic	Observed
	$E(Q)$	$\overline{Q}$ (95% C.I.)	$E(Q)$	$\overline{Q}$ (95% C.I.)	$E(Q)$	$\overline{Q}$ (95% C.I.)
0	9.00	9.08 (8.82,9.35)	9.00	9.06 (8.78,9.33)	9.00	9.28 (9.01,9.56)
0.05	13.50	13.52 (13.12,13.93)	12.61	12.58 (12.20,12.97)	9.94	10.09 (9.80,10.37)
0.10	18.00	17.73 (17.23,18.22)	16.22	16.44 (15.94,16.95)	10.89	10.92 (10.59,11.24)
0.15	22.50	22.72 (22.11,23.44)	19.83	19.12 (18.52,19.72)	11.83	11.72 (11.37,12.07)
0.20	27.00	27.34 (26.54,28.14)	23.44	22.80 (22.01,23.58)	12.78	12.43 (12.05,12.81)
0.30	36.00	36.61 (35.55,37.69)	30.67	29.48 (28.45,30.52)	14.67	14.84 (14.34,15.35)
0.40	45.00	44.31 (42.96,45.65)	37.89	37.74 (36.49,39.04)	16.56	16.62 (16.05,17.18)
0.50	54.00	53.56 (51.99,55.13)	45.11	44.99 (43.27,46.72)	18.44	18.46 (17.08,19.12)

W=Gradient of line of $E(Q)$ against σ_B^2

Figure 44: Power of the Q statistic against the between-study variance for different numbers of trials k when $\sum_{i=1}^{k} w_i$=100 and the weights are all equal



Key

· k=5, equal weight w=20

+ k=10, equal weight w=10

* k=20, equal weight w=5

1000 simulations at each point

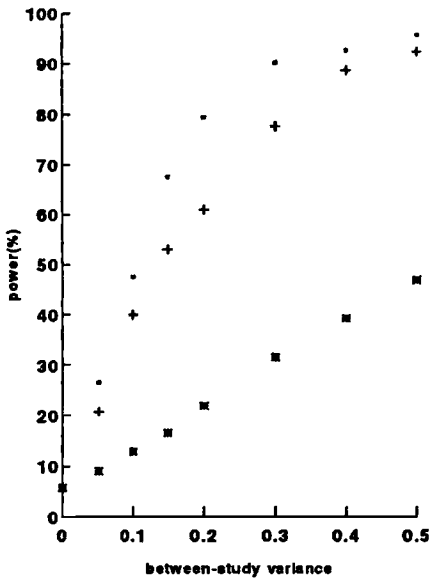
4.2.2 Power and the number of trials

It may further be deduced from equation (87) that the expectation of Q, under the assumption of homogeneity (i.e. when σ_B^2=0), increases as the number of trials in the meta-analysis increases. This increase in test statistic is, however, accompanied by an increase in the associated degrees of freedom. There are two ways of looking at the effect of the number of trials on the power, since changing the number of trials k in the meta-analysis necessarily implies that the weights change too. Firstly, if the weight allocated to each trial remains constant, w_i=10, i=1,...,k, in this example, as k increases (k=5, 10, and 20), so the total weight $\sum_{i=1}^{k} w_i$ will increase. On the other hand, if the total sum of weight $\sum_{i=1}^{k} w_i$ is kept constant, at 100 in this example, and the number of trials k is varied (k=5, 10, 20) then the individual weight for each trial w_i will change.

When w is kept constant, implying that $\sum_{i=1}^{k} w_i$ increases, the expectation of Q is larger and increases at a greater rate for larger k. However, since the degrees of freedom change as k changes, plots of $E(Q)$ against σ_B^2 are difficult to interpret in relation to the power of the test. From the simulations it can be seen that the power of the test also increases with increasing k (Figure 45) and, therefore, with increasing total information. Power of almost 100% is reached, for the example where k=20 (and $\sum_{i=1}^{k} w_i$=200) by the time σ_B^2 is equal to 0.3, while for the case where k=5 (and $\sum_{i=1}^{k} w_i$=50), the power only just reaches 65% at this point.

When $\sum_{i=1}^{k} w_i$ remains constant and the individual w vary with k, the larger the number of studies in the meta-analysis, the larger the value of $E(Q)$ for a given σ_B^2 and the steeper the gradient of the plot of $E(Q)$ against σ_B^2 (Figure 43). However, the gradients of these plots, for the examples chosen, are only slightly different. The corresponding power plots for each k (k=5, 10 and 20) are fairly similar to each other (Figure 44) and there is no one example which consistently has the greatest power over all values of σ_B^2. For small values of σ_B^2 the power is greatest in the case

Figure 45: Power of the Q statistic against the between-study variance for different numbers of trials k when each individual weight is equal to 10



Key

* k=5, $\sum_{i=1}^{k} w_i$=50

\+ k=10, $\sum_{i=1}^{k} w_i$=100

\- k=20, $\sum_{i=1}^{k} w_i$=200

1000 simulations at each point

where k=5, while for larger values the power is greatest in the case where k=20. In mid-range where a crossing over appears to take place the test statistic for the meta-analysis with 10 trials has greatest power. The confidence intervals for the simulated power values do, however, suggest that the power is not exactly the same for all three examples. Hence, it is difficult to summarise the relationship between the power of the test and the number of trials when the total information remains constant.

4.2.3 Power and weight

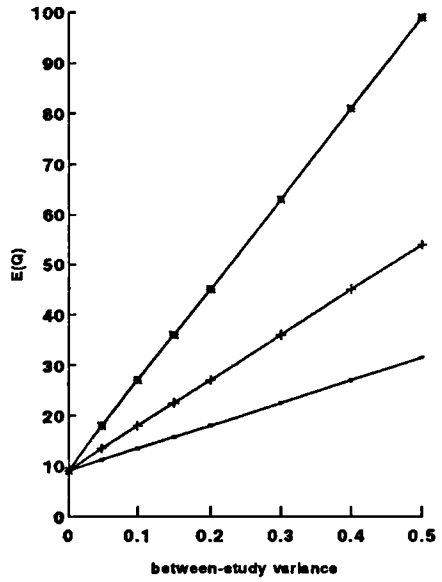
The most interesting, as well as most complex, aspect of this investigation is to assess the effect that changing the allocation of the weight has on the value of E(Q) and thus on power. On the simplest level, for a given k and when all weights are equal to one another, the expectation of the test statistic increases as the amount of total information increases (Figure 46). Hence, the larger the total information, the steeper the gradient of the plot of $E(Q)$ against σ_B^2 and the faster the rate of increase in $E(Q)$.

It may then be deduced, from (87), that as w_1 increases, E(Q) decreases (Figure 47). If $\sum_{i=1}^{k} w_i$ remains fixed, then for a given σ_B^2 and a given number of trials k, E(Q) is a maximum when $\sum_{i=1}^{k} w_i^2$ is a minimum. Hence, by minimising $\sum_{i=1}^{k} w_i^2$ under the constraint that $\sum_{i=1}^{k} w_i$ is constant, it is found that the maximum $E(Q)$ is obtained when the weight allocated to each study in the meta-analysis is the same. The maximum of the expected value of Q for a given $\sum_{i=1}^{k} w_i$, σ_B^2 and k, is therefore obtained by substituting $\sum_{i=1}^{k} w_i/k$ for w_i in (87),

$$(k-1) + \sigma_B^2 \frac{(k-1)}{k} \sum_{i=1}^{k} w_i \tag{89}$$

The minimum value of $E(Q)$, which is $(k-1)$, is approached as the gradient W tends to zero. That is as $\sum_{i=1}^{k} w_i^2 \rightarrow (\sum_{i=1}^{k} w_i)^2$, which occurs when one individual

Figure 46: Expectation of the Q statistic against the between-study variance for varying values of $\sum_{i=1}^{k} w_i$ when the number of trials k is 10 and the weights are all equal



Key

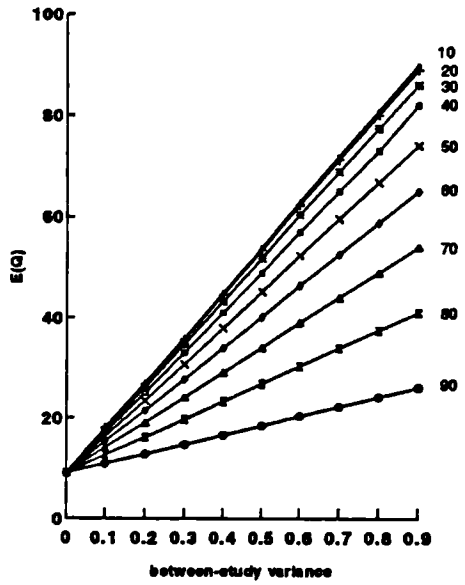
· $\sum_{i=1}^{k} w_i$=50

+ $\sum_{i=1}^{k} w_i$=100

* $\sum_{i=1}^{k} w_i$=200

1000 simulations at each point

Figure 47: Expectation of the Q statistic against the between-study variance for varying values of w_1 when $\sum_{i=1}^{k} w_i$=100 and the number of trials k is 10



1000 simulations at each point

w_i tends to $\sum_{i=1}^{k} w_i$. Hence, between the two extremes, the expected value of Q for a meta-analysis with given information decreases as the weights (or equivalently the variances) become more different (Figure 47). The same behaviour of $E(Q)$, that is decreasing $E(Q)$ with increasing differences in w_i, is observed for any value of $\sum_{i=1}^{k} w_i$, except that as the sum increases, the corresponding gradients become steeper.

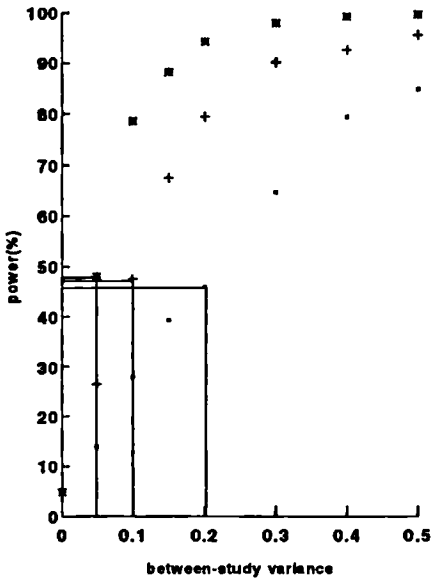
Considering the equivalent results for power, the meta-analyses with more information have generally greater power to detect heterogeneity. For a fixed number of trials, k=10 in these examples, the total amount of information included in a meta-analysis has a great effect on the power curve of the test, with greater power being achieved, for all values of σ_B^2, with increasing total information (Figure 48). When w_1 is varied for a fixed total amount of information, the power decreases as the weight given to this single trial increases (Figure 49). When $w_1 = 90$, the power is below 40% even when σ_B^2=0.5. This may not be surprising, since the within-study variances for trials 2 to 10 at 0.9 are still larger than the between-study variance. In contrast the power for the two other examples is over 90% when σ_B^2=0.5, and for the case where all weights are equal the between-study variance is 5 times that of the individual within-study variances.

4.2.4 Alternative ways of looking at power

Although the findings outlined in the previous three sections characterise the behaviour of the power of Q in relation to the factors of interest, it is perhaps difficult to form an idea of the practical implications. Hence, this section considers power in two alternative and more practically applicable ways. It can then be deduced for what practical situations the power of the test for heterogeneity is particularly low and therefore where extra caution may be required.

It is of practical relevance to investigate the power that the test for heterogene-

Figure 48: Power of the Q statistic against the between-study variance for varying values of $\sum_{i=1}^{k} w_i$ when the number of trials k is 10 and the weights are all equal



Key

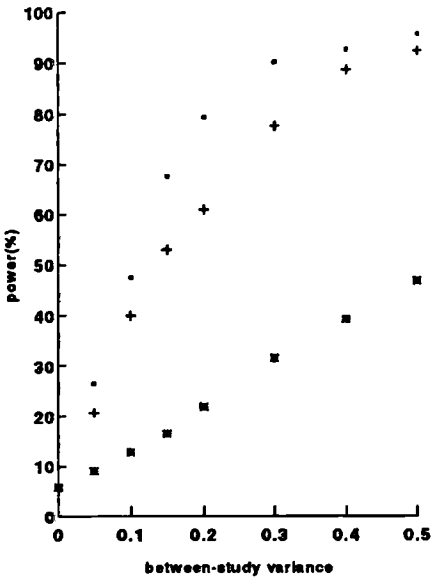
. $\sum_{i=1}^{k} w_i$=50, weight w=5

\+ $\sum_{i=1}^{k} w_i$=100, weight w=10

* $\sum_{i=1}^{k} w_i$=200, weight w=20

1000 simulations at each point

Figure 49: Power of the Q statistic against the between-study variance for varying values of w_1 when $\sum_{i=1}^{k} w_i$=100 and the number of trials k is 10



Key

. w_1=10

+ w_1=50

* w_1=90

1000 simulations at each point

ity has to detect a between-study variance σ_B^2 at least as large as the within-study variances v_i. If a meta-analysis is such that the v_i are much smaller than the σ_B^2 then the between-study component is of great importance, as its influence on the overall results (treatment effect estimate and confidence interval) will be substantial. On the other hand, the between-study variance is far less influential if it is small in comparison to the v_i. Considering the behaviour of $E(Q)$ for a between-study variance σ_B^2 equal in size to an 'average within-study variance' will provide an insight into the related power.

Firstly, assuming that all within-study variances are equal ($v_i = v, i = 1, ...k$), the between-study variance of interest is then simply equal to v. Hence, substituting v in equation (87) in place of σ_B^2 gives

$$E(Q) = (k-1) + v\left(\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i} \right) \tag{90}$$

But since equal variances imply equal weights and $w = 1/v$, equation (90) becomes

$$E(Q) = (k-1) + \frac{1}{w}\left(kw - \frac{kw^2}{kw} \right) = 2(k-1) \tag{91}$$

Hence, the expectation of Q when σ_B^2=v depends only on the number of trials included in the meta-analysis and is independent of the total weight. $E(Q)$ is, in fact, always twice the degrees of freedom in such a situation. Obviously as $\sum_{i=1}^{k} w_i$ changes, the value of v changes and so different values of σ_B^2 are being detected each time. As the number of studies in the meta-analysis increases, then the expectation of Q for σ_B^2=v also increases. Furthermore, the value $E(Q) = 2(k-1)$ is also the maximum value of $E(Q)$ that can be obtained when trying to detect the particular between-study variance σ_B^2=v, since the maximum always occurs when all trials receive the same weight (Section 4.2.3).

Looking at the results of power from this point of view, for a given number of trials k, is more complicated. The power depends not only on $E(Q)$, but also on the distribution of Q, and hence it cannot be stated that the power to detect a value of σ_B^2 equal to v remains constant for all values of $\sum_{i=1}^{k} w_i$. However, the power is likely to be similar for all values of $\sum_{i=1}^{k} w_i$ and, indeed, the simulations back this up. For k=10, the power was found to be between 45% and 50% for each $\sum_{i=1}^{k} w_i$ (Figure 48).

Alternatively, for a given $\sum_{i=1}^{k} w_i$, it can be stated that the meta-analysis with the largest number of trials has the greatest power because, although the power curves follow similar paths, the meta-analysis with the most studies has the largest within-study variance. In the example where $\sum_{i=1}^{k} w_i$=100, when k=20 the power to detect a between-study variance of v=0.2 is 70%, while the power to detect a between-study variance of v=0.05 when k=5 is only 30% (Figure 44).

The approach outlined above may be extended, still assuming that one wishes to detect the between-study variance equal to v of the previous example, by allowing the individual variances to be different. Hence, for a given $\sum_{i=1}^{k} w_i$, a between-study variance equal to the value of the within-study variance if all studies had the same variance and therefore the same weight $\sum_{i=1}^{k} w_i/k$ (i.e $\sigma_B^2 = 1/(\sum_{i=1}^{k} w_i/k)$ will be detected. Substituting $1/(\sum_{i=1}^{k} w_i/k)$ into (87) gives the required expectation,

$$E(Q) = (k-1) + \frac{k}{\sum_{i=1}^{k} w_i}\left(\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i} \right) \tag{92}$$

This can be simplified, and becomes

$$E(Q) = 2k - 1 - k\frac{\sum_{i=1}^{k} w_i^2}{(\sum_{i=1}^{k} w_i)^2} \tag{93}$$

Then, for a given k, where the total weight is allocated such that the percentage

weights ($(w_i/\sum_{j=1}^{k} w_j)$x100%) are the same, each trial weight w_i can be written as $x_i \sum_{j=1}^{k} w_j$ for any $\sum_{i=1}^{k} w_j$, where x_i is the fraction of the total weight taken by trial i and so $\sum_{i=1}^{k} x_i$=1. This implies that $\sum_{i=1}^{k} w_i^2/(\sum_{i=1}^{k} w_i)^2$ is equal to $\sum_{i=1}^{k} x_i^2$ and so is a constant, say c, for any $\sum_{i=1}^{k} w_i$ given a fixed percentagewise allocation of weights to k studies. Hence (93) becomes,

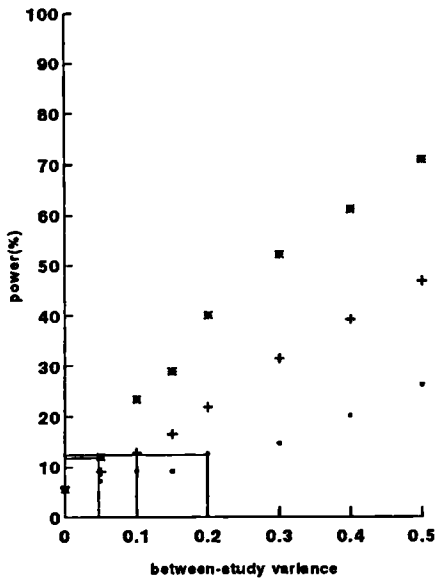
$$E(Q) = k(2 - c) - 1 \tag{94}$$

It can be seen from (94) that, where the weights are not all equal, $E(Q)$ depends not only on the number of trials k, but also on the percentage weights allocated to those k trials which determine the value of c. However, the expectation is still independent of the total information $\sum_{i=1}^{k} w_i$ and therefore provides a useful way of summarising the expectation of the test statistic since it eliminates the variables of σ_B^2 and $\sum_{i=1}^{k} w_i$. The dependence of $E(Q)$ on the total sum of weights is effectively removed by fixing the value of the between-study variance which it is deemed necessary to detect since σ_B^2 is derived from $\sum_{i=1}^{k} w_i$.

Looking at the corresponding results for the power of the test, it is found that the power to detect the particular value of $\sigma_B^2 = 1/(\sum_{i=1}^{k} w_i/k)$ for each value of w_1 (w_1=10, 50 and 90) is approximately constant for every value of $\sum_{i=1}^{k} w_i$. The power, at around 45–50%, being greatest when all weights are equal (Figure 48) and dropping to about 12% when $w_1 = 90$ (Figure 50). Alternatively, for a given $\sum_{i=1}^{k} w_i$ and a fixed percentage allocation of weights, there is an increase in power as k increases (Figure 44). However, the decreasing power with increasing w_1 is still evident for any allocation (Figures 44 and 51).

Hence, the simulations back up the analytical findings that, for a given number of trials in a meta-analysis, the power to detect a between-study variance as large as the within-study variance if all weights were equal, that is σ_B^2=$1/(\sum_{i=1}^{k} w_i/k)$,

Figure 50: Power of the Q statistic against the between-study variance for varying values of $\sum_{i=1}^{k} w_i$ when the number of trials k is 10 and w_1 takes 90% of the weight



Key

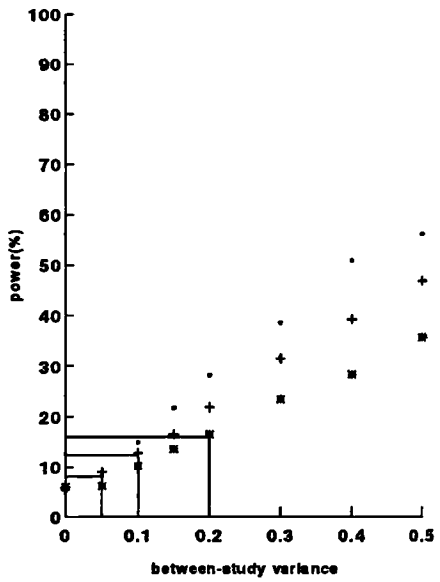
. $\sum_{i=1}^{k} w_i$=50

\+ $\sum_{i=1}^{k} w_i$=100

* $\sum_{i=1}^{k} w_i$=200

1000 simulations at each point

Figure 51: Power of the Q statistic against the between-study variance for varying numbers of trials k when $\sum_{i=1}^{k} w_i$=100 and w_1 takes 90% of the weight



Key

· k=5

+ k=10

* k=20

1000 simulations at each point

remains approximately constant for all values of $\sum_{i=1}^{k} w_i$.

The discussion here shows that for small k, and particularly when there is an uneven distribution of weight, the power of the test, defined in terms of being able to detect a between-study variance as large as the average within-study variance, is rather low. Hence, a second way of looking at power from a practical perspective, is to consider the 'effective sample size' necessary to maintain the same value of $E(Q)$ for a given σ_B^2, compared to the most powerful case, that of equal weighting. If k and σ_B^2 remain constant, then from (87) it may be seen that $E(Q)$ depends solely on the value of $W = \sum_{i=1}^{k} w_i - (\sum_{i=1}^{k} w_i^2 / \sum_{i=1}^{k} w_i)$. Then, for general $\sum_{i=1}^{k} w_i$, let W_e and T_e be the value of W and the value of $\sum_{i=1}^{k} w_i$ under the condition of equal weighting (w_i=w, for all i). Furthermore, assume that F is the multiplicative factor by which T_e must be increased in order that the same expectation as that achieved under equal weighting be maintained, and let each individual weight be written in terms of the percentage of the new total weight, that is $w_i = x_i F T_e$ where x_i is the fraction of the total weight taken by trial i. In order to maintain the same expectation with alternative weights, the new W must be equal to W_e. Hence, setting the expression for W obtained under the alternative unequal weighting, that is where w_i=$x_i F T_e$, equal to W_e gives

$$W_e = F T_e \sum_{i=1}^{k} x_i - \frac{F^2 T_e^2 \sum_{i=1}^{k} x_i^{\,2}}{F T_e \sum_{i=1}^{k} x_i} \tag{95}$$

where F is the factor to be calculated. Hence, rearranging (95) allows F to be found,

$$F = \frac{W_e}{T_e(1 - \sum_{i=1}^{k} x_i^2)} \tag{96}$$

As an example, let k=10, then for any value of $\sum_{i=1}^{k} w_i$ (i.e. total information or 'effective sample size'), $\sum_{i=1}^{k} w_i$=T_e obtained under equal weighting must be mul-

tiplied by the factors given in Table 40 in order to maintain the same value of $E(Q)$. Hence, when trial 1 takes 50% of the total weight, the effective sample size must be increased by a factor of 1.25 and when it takes 70%, the effective sample size must be doubled.

Table 40: Multiplicative factors for the 'effective sample sizes' required to maintain value of $E(Q)$ equal to that obtained under equal weighting for any total weight, where k=10 and σ_B^2 is fixed

Percentage weight to trial 1	'Sample size' required to maintain $E(Q)$
10	1
30	1.05
50	1.25
70	2.00
90	4.76

The diuretics trials data may be used as a practical illustration of this idea, but where the procedure works in reverse. The total weight under the observed unequal weighting scheme, T_o say, is known and the total weight under equal weighting T_e required to maintain the observed value of $E(Q)$ may be calculated. Obviously in this case T_e will be less than the observed total T_o. Again, since k and σ_B^2 are fixed, the interest lies in the value of W only. For an equal weighting situation, the value of W given in (95) may be simplified to $(T_e(k-1))/k$ since $x_i^2 = (1/k)^2$. Then, since T_e can be written in terms of T_o, that is as FT_o, this becomes

$$\frac{FT_o(k-1)}{k} \tag{97}$$

Setting (97) equal to the observed W, W_o say, which must be maintained, the factor

F can be found using

$$F = \frac{kW_o}{T_o(k-1)} \tag{98}$$

The within-study variances v_i in the diuretics trials meta-analysis vary quite considerably from 0.014 to 0.686, with $\hat{\sigma}_B^2$ equal to 0.23. The total observed information T_o is 125. If the weights were equal, then using (98) to calculate F, the same value of Q as that observed would be obtained with 73.6% of the information T_o actually observed. This means that a total weight of 92 (w=10.22) rather than 125 is required and hence, it can be seen that there is a considerable drop from the maximum power due to the unequal weighting of the trials and the test is less powerful.

4.3 Alternative Statistic for the Test of Heterogeneity

Due to the recognised low power of the test for heterogeneity using Q, an alternative statistic Q' has been proposed by Ewertz, Duffy et al. [110]. It includes a 'correction' which allows for the correlation between each individual study estimate $\hat{\theta}_i$ and the overall estimate $\hat{\theta}$. The idea behind the statistic Q' is that the 'correction' will cause the power of the test for heterogeneity to be increased. The statistic is given by

$$Q' = \sum_{i=1}^{k} \frac{(\hat{\theta}_i - \hat{\theta})^2}{(v_i - (\sum_{i=1}^{k} w_i)^{-1})} \tag{99}$$

It was stated that, under homogeneity, Q' has a chi-squared distribution on $(k-1)$ degrees of freedom [110], that is it has the same null distribution as the test statistic Q. This claim was checked using the basic simulation programs used in the previous applications. The distribution of the statistic Q' was obtained under the null hypothesis of homogeneity, and this was compared with both the χ_k^2 and the χ_{k-1}^2 distributions and the results are given in Section 4.3.1. Furthermore, in Section 4.3.2

the results obtained for Q from the simulations are compared with those obtained for Q'. Values of Q' and the associated power were obtained for a selection of examples used in the previous investigation of Q.

4.3.1 Distribution of Q'

Initially, the null distribution, that is the distribution under the homogeneous fixed effect model, of the test statistic was investigated. The type I error rates α (α=0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5) were obtained for Q' from the 1000 computer simulated data sets, where $\sum_{i=1}^{k} w_i$=100 and k=10. The process of obtaining α was carried out assuming a χ^2 distribution on both k=10 and $(k-1)$=9 degrees of freedom. The results (Table 41) show that for all three situations considered (w_1=10, 50 and 90) the values of α obtained are substantially larger than the theoretical values for a χ^2_{k-1} distribution. The values of α obtained assuming a χ^2_k distribution are also slightly larger then the theoretical values in all but one instance. Hence, these results suggest that the statistic Q' does not have a χ^2_{k-1} distribution, as was suggested. The distribution of the statistic would appear to be closer to a χ^2_k distribution, although this appears to be only approximate. As w_1 increases, the discrepancy between the theoretical value of α and that observed increases, indicating that the approximation to a χ^2_k distribution may be better when the weights are equal than when they are very different.

The distribution of Q', for the case where w_i=w for all i, was investigated further by means of chi-squared quantile plots of the simulated distribution of the statistic Q'. The points plotted were $(q_i, Q'_{(i)})$, where $Q'_{(i)}$ is the i^{th} smallest value of Q' and

$$q_i = F^{-1}((i - 3/8)/(N + 1/4))$$

where F is the cumulative distribution function of the χ^2 distribution and N is the

Table 41: Distribution of the Q' test statistic for heterogeneity under the null hypothesis of a homogeneous fixed effect model with $\sum_{i=1}^{k} w_i$=100 and k=10

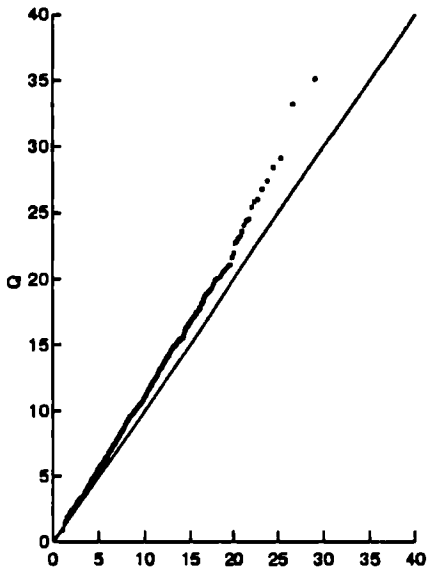
True α	Weight allocated to trial 1 w_1 ($w_2 = ... = w_{10}$)					
	10		50		90	
	$\chi^2_{9,\alpha}$	$\chi^2_{10,\alpha}$	$\chi^2_{9,\alpha}$	$\chi^2_{10,\alpha}$	$\chi^2_{9,\alpha}$	$\chi^2_{10,\alpha}$
0.01	0.018	0.014	0.024	0.016	0.023	0.017
0.05	0.085	0.059	0.097	0.061	0.102	0.074
0.10	0.162	0.110	0.168	0.120	0.177	0.128
0.20	0.280	0.216	0.280	0.225	0.293	0.236
0.30	0.381	0.311	0.388	0.310	0.423	0.326
0.40	0.512	0.401	0.491	0.401	0.524	0.434
0.50	0.601	0.516	0.574	0.495	0.611	0.533

number of values of the statistic, which in this case is 1000. If the statistic has a χ^2 distribution with the correct degrees of freedom, such a plot will be a straight line with a gradient of 1.

A χ^2_{k-1} and a χ^2_k quantile plot (Figures 52 and 53) were produced for the distribution of Q', using the values obtained from the simulation of the situation with equal weights (w_i=10). A χ^2_{k-1} plot for the usual test statistic Q was also produced (Figure 54), in order to have a standard plot by which to compare the plots of Q'. When the statistic Q is plotted against the quantiles of a χ^2_{k-1} distribution (Figure 54) the expected straight line is achieved. However, when Q' is plotted against the quantiles of a χ^2_{k-1} distribution, the straight line has a gradient which is steeper than 1 (Figure 52). The plot of Q' against the quantiles from a χ^2_k distribution (Figure 53), again produces a good straight line, and this time it is closer to a line with a gradient of 1. However, the gradient still appears to be slightly steeper than 1, with

a greater deviation being noticeable for larger values of q_i. Hence, when comparing this plot with that for Q, which does have a χ^2_{k-1} distribution, it suggests that the distribution of Q' is only approximately χ^2_k even when the weights are equal.

Figure 52: χ^2_{k-1} quantile plot for the distribution of Q'



4.3.2 Power of Q'

The power of the alternative test for heterogeneity using Q' would certainly be greater than that of Q, if it were compared to the χ^2_{k-1} distribution. However, it would appear that this is not the correct null distribution for the test statistic. Hence, for the simulations involving Q', the power was obtained under the assumption that the null distribution was approximately χ^2_k.

From a comparison of the results, it can be seen that in every case, the power of the test for heterogeneity using Q' is greater than that using Q (Table 42). This increase in power is, however, not very large in any of the three examples considered, being very slight when all weights are equal and getting larger as w_1 increases. It is

Figure 53: χ^2_k quantile plot for the distribution of Q'

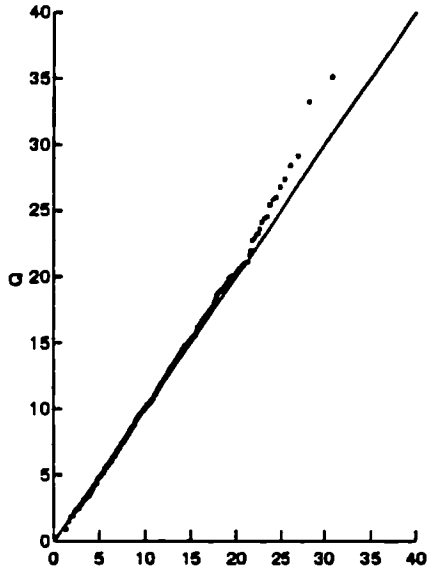


Figure 54: χ^2_{k-1} quantile plot for the distribution of Q

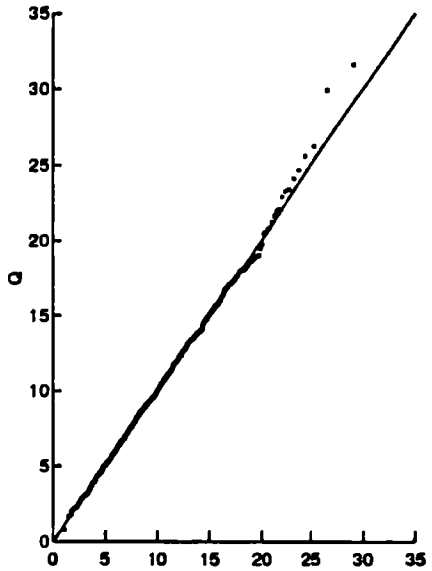


Table 42: Comparison of the power of the two statistics for heterogeneity, Q and Q', for three examples where $k = 10$ and $\sum_{i=1}^{k} w_i$=100

σ_B^2	Power (% significant from 1000 tests)					
	Weight allocated to trial 1 w_1 ($w_2 = ... = w_{10}$)					
	10		50		90	
	Q	Q'	Q	Q'	Q	Q'
0	5.3	5.9	5.2	6.1	5.9	7.4
0.05	26.5	28.5	20.7	25.0	9.0	11.8
0.10	47.5	50.1	39.9	43.2	12.8	19.2
0.15	67.5	68.9	53.0	57.1	16.5	22.4
0.20	79.4	80.6	60.9	65.1	21.9	28.0
0.30	90.2	91.2	77.6	80.0	31.5	40.0
0.40	92.7	93.6	88.7	90.2	39.2	46.1
0.50	95.7	96.2	92.4	93.2	46.8	52.8

only when w_1=90 that there is any materially useful improvement in power. However, in this case, the type I error rate (power when σ_B^2=0) is also increased above the 5% level, indicating the uncertainty about the null distribution. The standard errors for Q' are larger than those of Q, causing the 95% confidence intervals for Q' to be wider than those for Q. This indicates that the statistic Q' has a greater variability than Q.

It can be concluded from these simulations that the null distribution of the alternative statistic for heterogeneity is not χ^2_{k-1} and that the actual null distribution is only approximately χ^2_k. Furthermore, this approximation to the χ^2_k distribution gets less satisfactory as the weights become more uneven. Additional to this uncertainty regarding the null distribution of Q', little gain in power is achieved over Q. Hence, Q' cannot be recommended as a test of heterogeneity and so Q is still to be preferred.

4.4 Conclusions

The simulations in Section 4.2 have shown that the power of the test for heterogeneity using the test statistic Q will have low power in many practical situations. In particular the power will be less when the total amount of information available is small, either because k is small or because the individual trial estimates lack precision. Furthermore, as the weighting becomes uneven, the power drops from the maximum achievable for the total weight observed. Hence, in practice, care should be taken in the interpretation of a non-significant result from Q, especially if the meta-analysis has low observed total information or a highly uneven distribution of weight.

The simulated examples (Section 4.2) were chosen so that the behaviour of Q was investigated reasonably fully. However, further examples, such as ones in which all the weights are allowed to be different, may be required to obtain a complete picture. Furthermore, the results obtained only apply to detecting heterogeneity of a

specific form, that is heterogeneity that follows a normally distributed random effects model. Hence, although the expectation of Q is the same for a given σ_B^2 irrespective of the form of the heterogeneity, the results for power cannot necessarily be generalised to all situations. The power is dependent on the distribution of Q, which will be different under different alternative models. Hence, further work would be required to look at the power of the test under alternative heterogeneous models, especially as the random effects model may, in practice, be rather unrealistic.

It was shown in Section 4.3 that an alternative statistic Q', claimed to be more powerful than Q, in fact offers no improvement, particularly due to the uncertainty over the null distribution. However, it may be that other test statistics provide an improvement over Q. Indeed, based on the results of a simulation study, Jones et al. [109] recommend the use of the Breslow and Day statistic [111], which is similar to Q, but based on the Mantel-Haenszel estimate of overall odds ratio and used by StatXact [87], for situations with non-sparse data.

The results obtained in this chapter refer to the theoretical situation in which v_i are known. Hence, they cannot strictly be applied to the practical case, although they may be good approximations. The effect that estimating the v_i has on $E(Q)$ and the power of the test is addressed in the next chapter for the case of a continuous outcome measure.

5 The Effect of Estimated Weights on the Results of a Meta-Analysis

In all standard meta-analysis methods, it is assumed throughout that the weights are known. However, in practice the weights given by $w_i=1/v_i$ are, of course, estimated from the data. This chapter investigates the effect that the estimation of the weights has on the results obtained by the standard inverse-variance fixed effect (Section 1.5.1) and the standard random effects (Section 1.7.1) meta-analysis methods. The influence on the results of prime practical importance, that is the fixed effect and the random effects estimates of the overall treatment effect and their variances, are considered. In addition, the test for heterogeneity using Q and the estimate of the between-study variance are investigated. This work involved the use of quantitative data, as progress could be made analytically for this case. Computer simulations, similar to those used to investigate the power of the test for heterogeneity (Chapter 4) but with weights estimated from individually generated trial data, were programmed using FORTRAN (Section 3.3.1). Novel analytic methods were also pursued to try to obtain improved estimates allowing for the estimation of the weights. While exact analytic results proved difficult, approximations were obtained.

Section 5.1 describes the simulation methods and the theory behind the analytic work. The results from the simulations comparing both the standard estimates and the alternative estimates are then described in Section 5.2 and conclusions are drawn in Section 5.3.

5.1 Methods

The simulation procedure used is described in Section 5.1.1 and then the analytic theory is introduced in Section 5.1.2.

5.1.1 Simulation methods

The model used for the simulations in this chapter is described by (83) and (84) of Section 3.3.1 so that n_i observations are generated within each of the k trials. Then $\hat{\theta}_i$, which is the mean of y_{il}, l=1,...,n_i, in trial i in this case $\overline{y}_{i.}$, has a $N(\theta_i, v_i)$ distribution, where v_i=σ_i^2/n_i=$1/w_i$. The computer generated data y_{il}, $l = 1, ..., n_i$ and $i = 1, ..., k$, is used to calculate the individual trial estimate of treatment effect $\overline{y}_{i.}$ and its variance $\hat{v}_i$. This means that the estimated weight, denoted by $\hat{w}_i$ in this chapter and equal to $1/\hat{v}_i$, can be found. This reproduces the more realistic situation, where the estimated weights $\hat{w}_i$ are used to calculate the various statistics required. However, in practice, the measure of treatment effect $\hat{\theta}_i$ would often be a difference in means between two treatment groups, that is $\overline{y}_{i1.} - \overline{y}_{i2.}$, where $\overline{y}_{i1.}$=mean in the treatment group and $\overline{y}_{i2.}$=mean in the control group. Hence, assuming that the variance σ_i^2 is the same in both groups then $var(\hat{\theta}_i) = \frac{(n_{i1}+n_{i2})\sigma_i^2}{n_{i1}n_{i2}}$ and so the n_i in the simulations is equivalent to $(\frac{n_{i1}n_{i2}}{n_{i1}+n_{i2}})$ here. Therefore, the situation simulated is still a simplification of what usually occurs in reality, since only one group of observations, rather than two, is generated, but this resembles the situation where the reduction in blood pressure in an individual treatment group in the mild hypertension trial was considered (Section 3.4.1).

The number of observations n_i was taken to be the same in each trial, n say, and two different values of n were investigated. Firstly, n was set equal to 50, thus allowing each w_i to be estimated reasonably precisely. To provide a contrasting and extreme example n was then set equal to 5. The values of σ_B^2 (σ_B^2=0, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5) used when investigating the power of the test (Chapter 4) were again used, and 1000 repetitions were carried out at each point. The number of trials k in these simulations was fixed at 10 in order to keep things as simple as possible. Furthermore, the total information in the meta-analysis, $\sum_{i=1}^{k} w_i$, was fixed at 100. The examples from Chapter 4 where w_1 was allowed to vary (i.e. w_1=10, 50 and 90)

were repeated. For each repetition $\hat{\theta}_f$, $var(\hat{\theta}_f)$, Q, $\hat{\sigma}^2_B$, $\hat{\theta}_r$ and $var(\hat{\theta}_r)$ were calculated using standard estimates. However, for this chapter only, the subscript $\hat{w}$ will be used to denote that the weights are estimated from the data when calculating the estimates as opposed to being known. The mean values from the 1000 simulated data sets of the estimates of interest were then obtained and are, therefore, denoted by $\overline{\hat{\theta}_{f\hat{w}}}$, $\overline{var(\hat{\theta}_{f\hat{w}})}$, $\overline{Q_{\hat{w}}}$, $\overline{\hat{\sigma}^2_{B\hat{w}}}$, $\overline{\hat{\theta}_{r\hat{w}}}$ and $\overline{var(\hat{\theta}_{r\hat{w}})}$.

It should be mentioned that negative values of $\hat{\sigma}^2_{B\hat{w}}$, which are meaningless, may be obtained from the simulations. Hence, in practice the estimate of the between-study variance is taken to be $max\{\hat{\sigma}^2_{B\hat{w}}, 0\}$. However, even when known weights are used, this leads to a bias in the estimate of σ^2_B, and so for the simulation of the mean of $\hat{\sigma}^2_{B\hat{w}}$ negative values are included in order that the only bias occurring is caused by the estimation of the weights. However, when the random effect estimates and their variances are calculated $max\{\hat{\sigma}^2_{B\hat{w}}, 0\}$ is used.

5.1.2 Analytic methods

Analytic methods for quantifying the effect of estimating the weights are now pursued. An exact result is obtained for the expectation of a single estimated weight for a fixed effect model $E(\hat{w}_i)$, and this is then used as the basis for further approximations. The fact that

$$\frac{(n_i - 1)\hat{\sigma}^2_i}{\sigma^2_i} \sim \chi^2_{n_i-1} \quad i = 1, ..., k \tag{100}$$

can be taken as a starting point, since by using the probability density transformation

$$f(y) = f(x) \left| \frac{dx}{dy} \right| \tag{101}$$

with $x = \frac{(n_i-1)\hat{\sigma}^2_i}{\sigma^2_i}$ and $y = \hat{w}_i = n_i/\hat{\sigma}^2_i$, the probability density function of $\hat{w}_i$ may be

obtained:

$$f(\hat{w}_i) = \frac{1}{\Gamma(\frac{n_i-1}{2})} \left[\frac{n_i(n_i-1)}{2\sigma_i^2} \right]^{(n_i-1)/2} exp \left\{ \frac{-n_i(n_i-1)}{2\sigma_i^2 \hat{w}_i} \right\} \frac{1}{\hat{w}_i^{(n_i-1)/2}} \qquad (102)$$

where $\Gamma(\alpha)$ is the gamma function for α, and for integral α, $\Gamma(\alpha)=(\alpha-1)!$. Hence, the expectation and the variance of $\hat{w}_i$ can be calculated by integration methods, and in fact, it can be shown that,

$$E(\hat{w}_i) = \int \hat{w}_i f(\hat{w}_i) d\hat{w}_i = \frac{(n_i-1)}{(n_i-3)} w_i \qquad (103)$$

and

$$var(\hat{w}_i) = E(\hat{w}_i^2) - \{E(\hat{w}_i)\}^2 = 2 \left(\frac{n_i-1}{n_i-3} \right)^2 \frac{{w_i}^2}{(n_i-5)} \qquad (104)$$

The result for $E(\hat{w}_i)$ can now be used to consider the effect that estimating the weights has on the expected values of the estimates and variances of the overall treatment effect. The variance of $\hat{w}_i$ (104), however, did not prove to be useful in this regard since the exact analytic results became too complicated. The result that $E(\hat{w}) = f_i w_i$ where $f_i = \frac{(n_i-1)}{(n_i-3)}$ is now utilised to provide approximate adjusted estimates for both fixed effect and random effects meta-analyses.

The notation to be used in this chapter is firstly explained. As defined above, a standard estimate using estimated weights is identified with the subscript $\hat{w}$, for example $\hat{\theta}_{f_{\hat{w}}}$. Furthermore, a subscript f, for example $\hat{\theta}_{f_f}$, denotes an approximate analytic result based on the approximation that the weight $\hat{w}_i$ is known and equal to $f_i w_i$, while a subscript a, for example $\hat{\theta}_{f_a}$, denotes an approximate estimate of the parameter which may be obtained in practice containing $\hat{w}_i$ and f_i, but not the unknown w_i.

Fixed effect model: From the result in (103), one approach to obtaining improved estimates is to assume that each estimated weight $\hat{w}_i$ is equal to $f_i w_i$. Hence, by dividing each estimated weight $\hat{w}_i$ calculated in practice from the data by f_i a weight is produced which is on average closer to the true value $w_i=1/v_i$. Hence, an adjusted estimate of the overall treatment effect from a fixed effect model is given by

$$\hat{\theta}_{f_a} = \frac{\sum_{i=1}^{k} \frac{\hat{w}_i}{f_i}\hat{\theta}_i}{\sum_{i=1}^{k} \frac{\hat{w}_i}{f_i}} \tag{105}$$

with a variance, which if obtained by simple substitution of $\hat{w}_i/f_i$ for w_i, is

$$var(\hat{\theta}_{f_a}) = \frac{1}{\sum_{i=1}^{k} \frac{\hat{w}_i}{f_i}} \tag{106}$$

However, in practice the fixed effect estimate of treatment effect is usually estimated by $\hat{\theta}_{f_{\hat{w}}} = \frac{\sum_{i=1}^{k} \hat{w}_i\hat{\theta}_i}{\sum_{i=1}^{k} \hat{w}_i}$ and its variance by $var(\hat{\theta}_{f_{\hat{w}}}) = 1/\sum_{i=1}^{k} \hat{w}_i$. This variance will not be the true variance of $\hat{\theta}_{f_{\hat{w}}}$ in practice since it does not allow for the extra variation caused by the estimation of the weights; it is really the variance of $\sum_{i=1}^{k} w_i\hat{\theta}_i/\sum_{i=1}^{k} w_i$ not $\hat{\theta}_{f_{\hat{w}}}$. By making the assumption that $\hat{w}_i$ is known and equal to $f_i w_i$, rather than w_i, an approximation to $\hat{\theta}_{f_{\hat{w}}}$ may be obtained and is given by

$$\hat{\theta}_{f_f} = \frac{\sum_{i=1}^{k} f_i w_i\hat{\theta}_i}{\sum_{i=1}^{k} f_i w_i} \tag{107}$$

The variance of $\hat{\theta}_{f_f}$ may then be derived and takes the form

$$var(\hat{\theta}_{f_f}) = \frac{\sum_{i=1}^{k} f_i^2 w_i}{(\sum_{i=1}^{k} f_i w_i)^2} \tag{108}$$

Then an approximate 'adjusted' variance of $\hat{\theta}_{f_{\hat{w}}}$ which may be calculated in practice, allowing, to some extent at least, for the estimation of the weights can be obtained

from (108) by replacing $f_i w_i$ by $\hat{w}_i$

$$var_a(\hat{\theta}_{f_f}) = \frac{\sum_{i=1}^{k} f_i \hat{w}_i}{(\sum_{i=1}^{k} \hat{w}_i)^2} \tag{109}$$

where $\hat{w}_i$ are the weights estimated from the data. The variance given in (109) can be considered as an approximation to the variance of $\hat{\theta}_{f_{\hat{w}}}$ too since $\hat{\theta}_{f_f} \simeq \hat{\theta}_{f_{\hat{w}}}$. It might, therefore, be anticipated that this expression will provide an improved estimate of the variance of the standard fixed effect estimate of treatment effect $\hat{\theta}_{f_{\hat{w}}}$ in a practical situation.

If the number of observations in each study is equal, that is if f_i=f for all i, then $\hat{\theta}_{f_{\hat{w}}}$ becomes equal to $\hat{\theta}_{f_a}$, the approximate adjusted variance $var_a(\hat{\theta}_{f_f})$ (109) simplifies to $f/\sum_{i=1}^{k} \hat{w}_i$ which is then equal to $var(\hat{\theta}_{f_a})$ (106). This variance appears to be sensible in so far as it will give a variance greater than $1/\sum_{i=1}^{k} \hat{w}_i$ since $f > 1$, thus reflecting additional uncertainty included because of the estimation of the w_i. However, in practice, using the standard methods, the variance of the overall treatment effect is found using $1/\sum_{i=1}^{k} \hat{w}_i \simeq 1/f\sum_{i=1}^{k} w_i$. This is obviously incorrect and the variance calculated in this way will in fact be too small since $1/f\sum_{i=1}^{k} w_i$ is even smaller than $1/\sum_{i=1}^{k} w_i$.

Random effects model: The case of the random effects model is more complicated than that of the fixed effect model since the between-study variance must be obtained before the overall treatment effect can be estimated. The estimate of the between-study variance is derived using the test statistic for heterogeneity Q. In practice, Q, denoted by $Q_{\hat{w}}$ in this chapter, is given by $\sum_{i=1}^{k} \hat{w}_i(\hat{\theta}_i - \hat{\theta}_{f_{\hat{w}}})^2$. Hence, the expected value of $Q_{\hat{w}}$ must first be obtained and

$$E(Q_{\hat{w}}) = E(\sum_{i=1}^{k} \hat{w}_i(\hat{\theta}_i - \theta)^2) - E((\sum_{i=1}^{k} \hat{w}_i)(\hat{\theta}_{f_{\hat{w}}} - \theta)^2) \tag{110}$$

The covariance term of $\hat{w}_i$ and $(\hat{\theta}_{f_{\hat{w}}} - \theta)^2$ will not be zero since the terms are not independent owing to the fact that $\hat{\theta}_{f_{\hat{w}}}$ involves $\hat{w}_i$. An evaluation of this covariance term is difficult, and hence approximate results are again found by making the assumption as before, that is that the true weights are known and can be obtained by dividing each $\hat{w}_i$ by f_i. Then calculating the test statistic, say Q_a, gives

$$Q_a = \sum_{i=1}^{k} \frac{\hat{w}_i}{f_i}(\hat{\theta}_i - \hat{\theta}_{f_a})^2 \tag{111}$$

which has an expectation analogous to (12),

$$E(Q_a) = (k-1) + \sigma_B^2 \left(\sum_{i=1}^{k} \frac{\hat{w}_i}{f_i} - \frac{\sum_{i=1}^{k} \hat{w}_i^2/f_i^2}{\sum_{i=1}^{k} \hat{w}_i/f_i} \right) \tag{112}$$

Then assuming the weights are known, using the method of moments and equating $E(Q_a)$ (112) with Q_a and rearranging, in the manner for deriving the D&L estimate of the between-study variance with known weights (Section 1.7.1), one expression for the approximate adjusted value of the between-study variance $\hat{\sigma}^2_{B a_1}$ may be obtained:

$$\hat{\sigma}^2_{B a_1} = \frac{Q_a - (k-1)}{\left(\sum_{i=1}^{k} \frac{\hat{w}_i}{f_i} - \frac{\sum_{i=1}^{k} \hat{w}_i^2/f_i^2}{\sum_{i=1}^{k} \hat{w}_i/f_i} \right)} \tag{113}$$

Alternatively, the expectation of the approximate test statistic $Q_f = \sum_{i=1}^{k} f_i w_i (\hat{\theta}_i - \hat{\theta}_{f_f})^2$, which is $Q_{\hat{w}}$ with $\hat{w}_i$ replaced by $f_i w_i$ may be written as

$$E(Q_f) = E(\sum_{i=1}^{k} f_i w_i (\hat{\theta}_i - \hat{\theta}_{f_f})^2) = \sum_{i=1}^{k} f_i w_i var(\hat{\theta}_i) - \sum_{i=1}^{k} f_i w_i var(\hat{\theta}_{f_f}) \tag{114}$$

Expressing the variances in terms of w_i and σ_B^2 produces the following formula

$$E(Q_f) = \left(\sum_{i=1}^{k} f_i - \frac{\sum_{i=1}^{k} f_i^2 w_i}{\sum_{i=1}^{k} f_i w_i} \right) + \sigma_B^2 \left(\sum_{i=1}^{k} f_i w_i - \frac{\sum_{i=1}^{k} f_i^2 w_i^2}{\sum_{i=1}^{k} f_i w_i} \right) \quad (115)$$

Using the method of moments produces an estimate of σ_B^2 expressed in terms of w_i and f_i,

$$\hat{\sigma}_{Bf}^2 = \frac{Q_f - \left(\sum_{i=1}^{k} f_i - \frac{\sum_{i=1}^{k} f_i^2 w_i}{\sum_{i=1}^{k} f_i w_i} \right)}{\sum_{i=1}^{k} f_i w_i - \frac{\sum_{i=1}^{k} f_i^2 w_i^2}{\sum_{i=1}^{k} f_i w_i}} \quad (116)$$

This can be rewritten as follows, by substituting $\hat{w}_i$ for $f_i w_i$, so that an adjusted estimate of σ_B^2, $\hat{\sigma}_{Ba_2}^2$ say, is obtained,

$$\hat{\sigma}_{Ba_2}^2 = \frac{Q_{\hat{w}} - \left(\sum_{i=1}^{k} f_i - \frac{\sum_{i=1}^{k} f_i \hat{w}_i}{\sum_{i=1}^{k} \hat{w}_i} \right)}{\sum_{i=1}^{k} \hat{w}_i - \frac{\sum_{i=1}^{k} \hat{w}_i^2}{\sum_{i=1}^{k} \hat{w}_i}} \quad (117)$$

However, using the standard D&L estimator, the between-study variance calculated in practice is actually given by

$$\hat{\sigma}_{B\hat{w}}^2 = \frac{\sum_{i=1}^{k} \hat{w}_i (\hat{\theta}_i - \hat{\theta}_{f_{\hat{w}}})^2 - (k-1)}{\sum_{i=1}^{k} \hat{w}_i - \frac{\sum_{i=1}^{k} \hat{w}_i^2}{\sum_{i=1}^{k} \hat{w}_i}} \quad (118)$$

which, assuming that $\hat{w}_i = f_i w_i$, is approximately equivalent to

$$\frac{Q_f - (k-1)}{\sum_{i=1}^{k} f_i w_i - \frac{\sum_{i=1}^{k} f_i^2 w_i^2}{\sum_{i=1}^{k} f_i w_i}} \quad (119)$$

This expression is a biased estimate of σ_B^2.

If $n_i = n$ for all i, then from (111) $Q_a = Q_{\hat{w}}/f$, where $f = (n-1)/(n-3)$. The estimate of the between-study variance $\hat{\sigma}^2_{Ba_1}$ (113) becomes

$$\hat{\sigma}^2_{Ba_1} = \frac{(Q_{\hat{w}}/f) - (k-1)}{\frac{1}{f}\left(\sum_{i=1}^{k} \hat{w}_i - \frac{\sum_{i=1}^{k} \hat{w}_i^2}{\sum_{i=1}^{k} \hat{w}_i}\right)} \tag{120}$$

which is equal to $\hat{\sigma}^2_{Ba_2}$ (117) when $f_i = f$ for all i. Hence, for equal n_i the adjusted between-study variance estimator can be denoted by $\hat{\sigma}^2_{Ba}$.

Consideration of the random effects estimate and variance is problematic since even the expectation of a single estimated weight $\hat{w}_i^* = 1/(\hat{v}_i + \hat{\sigma}^2_{B\hat{w}})$ is complicated to obtain. The biased estimate of the between-study variance $\hat{\sigma}^2_{B\hat{w}}$ usually obtained in practice could be replaced by one of the adjusted estimates $\hat{\sigma}^2_{Ba_1}$, $\hat{\sigma}^2_{Ba_2}$ or $\hat{\sigma}^2_{Ba}$. Then the random effects estimate of the overall treatment effect would be given by

$$\hat{\theta}_{ra} = \frac{\sum_{i=1}^{k} \frac{\hat{\theta}_i}{(\hat{v}_i + \hat{\sigma}^2_{Ba})}}{\sum_{i=1}^{k} \frac{1}{(\hat{v}_i + \hat{\sigma}^2_{Ba})}} \tag{121}$$

Furthermore, by simple substitution

$$var(\hat{\theta}_{ra}) = \frac{1}{\sum_{i=1}^{k} \frac{1}{(\hat{v}_i + \hat{\sigma}^2_{Ba})^2}} \tag{122}$$

whereas in practice, the variance is calculated by $1/\sum_{i=1}^{k} \hat{w}_i^* = 1/\sum_{i=1}^{k} 1/(\hat{v}_i + \hat{\sigma}^2_{B\hat{w}})$, which is incorrect. However, it is not clear in which direction the bias will be since there is the effect of estimating both $\hat{v}_i$ and $\hat{\sigma}^2_{B\hat{w}}$ to consider.

The approximate 'adjusted' estimates, $var_a(\hat{\theta}_{fj})$ (109), Q_a (111) and $\hat{\sigma}^2_{Ba}$ (120) for cases where n_i are all equal, since this is the simplest situation and in which $\hat{\theta}_{f\hat{w}} = \hat{\theta}_{fa}$, $var_a(\hat{\theta}_{fj}) = var(\hat{\theta}_{fa})$ and $\hat{\sigma}^2_{Ba_1} = \hat{\sigma}^2_{Ba_2}$, are compared with the equivalent

standard estimates in the next section. The simulations also provide an idea of the extent of the bias of the standard estimates.

5.2 Results

The results from the simulations for the fixed effect model are described and discussed in Section 5.2.1. Similarly, Section 5.2.2 considers the results for Q and Section 5.2.3 those for σ_B^2. Finally the results for the random effects model are discussed in Section 5.2.4. For ease of reference a table of the notation used in this chapter is provided (Table 43).

Table 43: Table of notation for Chapter 5

Variable	Usual estimate with estimated weights	Approximate analytic result with $\hat{w}_i = f_i w_i$	Adjusted estimate with $\hat{w}_i$ & f_i	Mean from simulations
Fixed effect estimate	$\hat{\theta}_{f\bullet}$	$\hat{\theta}_{ff}$	$\hat{\theta}_{fa}$	$\overline{\hat{\theta}_{f\bullet}}$
Variance of fixed effect estimate	$var(\hat{\theta}_{f\bullet})$	$var(\hat{\theta}_{ff})$	$var_a(\hat{\theta}_{ff}) \simeq var_a(\hat{\theta}_{f\bullet})$	$\overline{var(\theta_{f\bullet})}$
Test statistic for heterogeneity	$Q_\bullet$	Q_f	Q_a	$\overline{Q}_\bullet$
Between-study variance	$\hat{\sigma}^2_{B\bullet}$	$\hat{\sigma}^2_{Bf}$	$\hat{\sigma}^2_{Ba}$	$\overline{\hat{\sigma}^2_{B\bullet}}$
Random effects estimate	$\hat{\theta}_{r\bullet}$	-	$\hat{\theta}_{ra}$	$\overline{\theta_{r\bullet}}$
Variance of random effects estimate	$var(\hat{\theta}_{r\bullet})$	-	-	$\overline{var(\theta_{r\bullet})}$

5.2.1 Fixed Effect Model

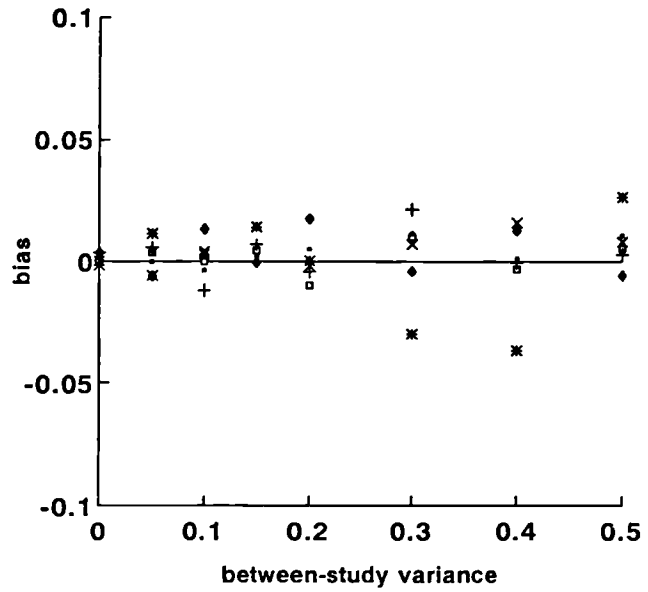
In this section $\hat{\theta}_{f_{\hat{w}}} = \sum_{i=1}^{k} \hat{w}_i\hat{\theta}/\sum_{i=1}^{k} \hat{w}_i$, that is the estimate of the overall treatment effect which is usually calculated in practice using weights calculated from the simulated data, is checked to make sure that it is unbiased. In the examples considered $\hat{\theta}_{f_{\hat{w}}} = \hat{\theta}_{f_a}$, since f_i=f for all i and so $\hat{\theta}_{f_a} = \frac{1}{f}\sum_{i=1}^{k} \hat{w}_i\hat{\theta}_i/\frac{1}{f}\sum_{i=1}^{k} \hat{w}_i = \hat{\theta}_{f_{\hat{w}}}$. Furthermore, two methods for calculating the variance of $\hat{\theta}_{f_{\hat{w}}}$, that is using the standard $1/\sum_{i=1}^{k} \hat{w}_i$ and also the adjusted variance $var_a(\hat{\theta}_{f_f}) = var_a(\hat{\theta}_{f_{\hat{w}}}) = f/\sum_{i=1}^{k} \hat{w}_i$ (109) proposed in Section 5.1.2. This latter variance is also equal to the variance of $\hat{\theta}_{f_a}$ (106) in the case where n_i=n for all i.

Initially, however, it is useful to check the behaviour of an individual estimated weight $\hat{w}_i$. When the number of observations in a study is large, f_i is very close to one and therefore the expectation of $\hat{w}_i$ is approximately equal to w_i. However, when the number of observations in a study is small, f_i is greater than one and in the extreme case when n_i=5, f_i=2. This characteristic may be illustrated using the values of w_1 obtained from the simulations. The mean value of $\hat{v}_1$ obtained from 1000 simulations when w_1=90 and n is 5 was 0.01151, which compares well with the true value of 0.01111. However, the mean value of $\hat{w}_1$ was 169.20, which is reasonably close to the expected value of 180, that is $2w_1$, and is certainly much larger than the value of w_1=90. Hence the simulations back up the analytic findings that the weights are, on average, inflated by the estimation of the variances v_i.

The results from the simulations indicate that there is no systematic bias (Figure 55) caused by the estimation of the weights in the overall fixed effect estimate of treatment effect. The plot displays $(\overline{\hat{\theta}_{f_{\hat{w}}}} - \theta)$, where $\overline{\hat{\theta}_{f_{\hat{w}}}}$ $(=\overline{\hat{\theta}_{f_a}})$ is the mean value of the estimates of θ from the 1000 simulated sets of data, and it can be seen that the points are randomly distributed either side of zero, with the variation becoming greater as σ_B^2 increases. The fact that no bias is seen is not surprising, since the estimate is still a weighted average of $\hat{\theta}_i$, but with weights other than the true w_i.

However, it is with regards to variance that the estimation of the weights is likely to have an impact, since the standard variance does not account for the estimation of the weights.

Figure 55: Plot showing the bias in the fixed effect estimate of the overall treatment effect $(\overline{\hat{\theta}_{f_{\hat{w}}}} - \theta)$ against the between-study variance



Key

. w_1=10, n=50 ▫ w_1=10, n=5

\+ w_1=50, n=50 × w_1=50, n=5

* w_1=90, n=50 ◆ w_1=90, n=5

1000 simulations at each point

The average simulated variance $\overline{var(\hat{\theta}_{f_{\hat{w}}})}$, obtained using $1/\sum_{i=1}^{k} \hat{w}_i$, is smaller than the theoretical analytic variance obtained under the assumption that all weights are known and equal to w_i, that is $1/\sum_{i=1}^{k} w_i$=0.01, in all three examples (w_1=10, 50 and 90) (Table 44). This is as expected since $1/\sum_{i-1}^{k} \hat{w}_i$ calculated in practice is approximately equal to $1/\sum_{i=1}^{k} f_i w_i$, or $1/f\sum_{i=1}^{k} w_i$ when $f_i = f$ for all i, and f_i is always greater than 1 (Section 5.1.2). In other words the variance will be calculated

using the estimated weights which are, on average, larger than the theoretical weights. Hence, the reciprocal of the sum of the estimated weights will tend to be smaller than the reciprocal of the sum of the true weights. The variance when n=5 is smaller than that when n=50. Hence, the simulated results are consistent with the analytic finding that the weights are inflated to a greater extent in the former situation due to the larger value of f. When n=5 the bias is clearly dependent on the allocation of the weight, with a greater discrepancy occurring for an uneven allocation of weight (Table 44). However, the bias is not dependent on the value of σ_B^2.

Table 44: Standard estimated variance of the fixed effect estimate when $1/\sum_{i=1}^{k} w_i$=0.01 for different allocations of weight

Between-study variance (σ_B^2)	Mean from simulations ($\overline{var(\hat{\theta}_{f_{\hat{w}}})}$)					
	number of observations in each trial (n)					
	n=50			n=5		
	w_1=10	w_1=50	w_1=90	w_1=10	w_1=50	w_1=90
0.00	0.00963	0.00969	0.00986	0.00596	0.00660	0.00885
0.05	0.00963	0.00969	0.00993	0.00604	0.00641	0.00855
0.10	0.00960	0.00967	0.00991	0.00607	0.00651	0.00884
0.15	0.00958	0.00968	0.01008	0.00597	0.00654	0.00859
0.20	0.00965	0.00967	0.00985	0.00595	0.00650	0.00878
0.30	0.00962	0.00969	0.01000	0.00597	0.00653	0.00861
0.40	0.00963	0.00972	0.00991	0.00590	0.00654	0.00855
0.50	0.00966	0.00967	0.00994	0.00606	0.00661	0.00879

w_1=Weight allocated to trial 1

The true variance of $\hat{\theta}_{f_{\hat{w}}}$ will, however, be larger even than $1/\sum_{i=1}^{k} w_i$ in practice, since the estimation of the weights introduces some additional variation. Hence, the actual variance of the simulated means, that is $\hat{var}(\hat{\theta}_{f_{\hat{w}}})$, as opposed to the mean

Table 45: Comparison of the mean standard estimated variance and the observed variance of the fixed effect estimate when $1/\sum_{i=1}^{k} w_i$=0.01, for different allocations of weight under homogeneity (i.e. σ_B^2=0)

Weight given to trial 1 (w_1)	Number of observations in each trial (n)			
	50		5	
	$\overline{var(\hat{\theta}_{f_{\hat{w}}})}$	$\hat{var}(\hat{\theta}_{f_{\hat{w}}})$	$\overline{var(\hat{\theta}_{f_{\hat{w}}})}$	$\hat{var}(\hat{\theta}_{f_{\hat{w}}})$
10	0.00963	0.01014	0.00596	0.01885
50	0.00969	0.01122	0.00660	0.01957
90	0.00986	0.00935	0.00885	0.01751

of the 1000 simulated variances $\overline{var(\hat{\theta}_{f_{\hat{w}}})}$, was calculated in order to obtain an estimate of $var(\hat{\theta}_{f_{\hat{w}}})$ allowing for the estimation of the weights. The variances $\hat{var}(\hat{\theta}_{f_{\hat{w}}})$, obtained under the assumption of homogeneity (σ_B^2=0), were generally found to be greater than $1/\sum_{i=1}^{k} w_i$=0.01 (Table 45), and the increases were greater when n=5 than n=50. Since under the fixed effect model the variance of $\hat{\theta}_{f_{\hat{w}}}$ is calculated assuming homogeneity, whatever the value of the true σ_B^2, the mean variance should be equal to that when σ_B^2=0. Hence, the mean variances for all values of σ_B^2 given in Table 44 can be compared with $\hat{var}(\hat{\theta}_{f_{\hat{w}}})$ in Table 45. For all three values of w_1 it can be seen that the standard estimate of the variance will always be too small.

The alternative approximate expression derived for the variance of $\hat{\theta}_{f_{\hat{w}}}$, that is $var_a(\hat{\theta}_{f_{\hat{w}}}) = f/\sum_{i=1}^{k} \hat{w}_i$, was also used to calculate the variance for each repetition in each simulation example, and the mean of these was obtained. The results for the mean of the adjusted variance $\overline{var_a(\hat{\theta}_{f_{\hat{w}}})}$ (Table 46) can then be compared to $\overline{var(\hat{\theta}_{f_{\hat{w}}})}$ obtained using $1/\sum_{i=1}^{k} \hat{w}_i$ (Table 44) and also to the true variances obtained from the 1000 simulated values of $\hat{\theta}_{f_{\hat{w}}}$ (Table 45). For n=5 a clear improvement in the estimation of $var(\hat{\theta}_{f_{\hat{w}}})$ is shown when using $f/\sum_{i=1}^{k} \hat{w}_i$ as opposed to $1/\sum_{i=1}^{k} \hat{w}_i$, particularly for large w_1. Where w_1=90, $\hat{var}(\hat{\theta}_{f_{\hat{w}}})$=0.01751 and all the mean adjusted

variances $\overline{var_a(\hat{\theta}_{f_{\hat{w}}})}$ agree with this to three decimal places (Table 46). However, for the equal weighting case, the adjustment is less good, although the adjusted mean variances have at least increased from the unadjusted ones (Table 44). For n=50, the error in the unadjusted variances are small anyway (Table 44), but the adjusted variance still causes the mean variance to increase towards $\hat{var}(\hat{\theta}_{f_{\hat{w}}})$. However, when w_1 was equal to 90 (Table 45), $\hat{var}(\hat{\theta}_{f_{\hat{w}}})$ was, surprisingly, smaller then $1/\sum_{i=1}^{k} w_i$=0.01, although the difference was small enough to be due to sampling error.

Table 46: Alternative estimated variance of the fixed effect estimate when $1/\sum_{i=1}^{k} w_i$=0.01 for different allocations of weight

Between-study variance (σ_B^2)	Mean from simulations ($\overline{var_a(\hat{\theta}_{f_{\hat{w}}})}$)					
	number of observations in each trial (n)					
	n=50			n=5		
	w_1=10	w_1=50	w_1=90	w_1=10	w_1=50	w_1=90
0.00	0.01004	0.01011	0.01028	0.01193	0.01320	0.01770
0.05	0.01004	0.01010	0.01035	0.01208	0.01282	0.01710
0.10	0.01001	0.01009	0.01034	0.01214	0.01301	0.01769
0.15	0.00998	0.01009	0.01050	0.01194	0.01309	0.01718
0.20	0.01006	0.01008	0.01027	0.01190	0.01299	0.01757
0.30	0.01003	0.01010	0.01043	0.01194	0.01306	0.01723
0.40	0.01004	0.01013	0.01033	0.01181	0.01308	0.01710
0.50	0.01007	0.01008	0.01036	0.01213	0.01321	0.01757

w_1=Weight allocated to trial 1

It is clear from the simulations that when calculating a variance for an overall fixed effect estimate in a practical situation, using the standard estimate $1/\sum_{i=1}^{k} \hat{w}_i$ is likely to produce a value which is too small. Using $\sum_{i=1}^{k} f_i\hat{w}_i/(\sum_{i=1}^{k} \hat{w}_i)^2$ (or $f/\sum_{i=1}^{k} \hat{w}_i$ in the case where the n_i are all equal) appears generally to be a better

alternative, producing a value which is larger than $1/\sum_{i=1}^{k} w_i$.

5.2.2 Test statistic for heterogeneity

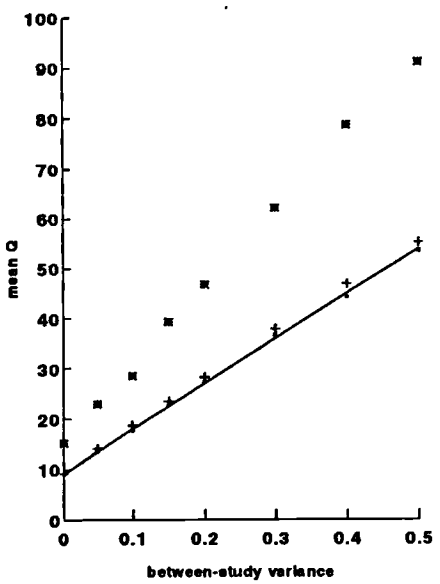
In this section $Q_{\hat{w}}$, the test statistic for heterogeneity obtained in practice, is compared to the theoretical value of $E(Q)$ obtained under the assumption that the weights are known. Then since the simulations are of examples where n_i=n for all i, an approximate adjusted estimate of the true Q may be obtained by simply dividing $Q_{\hat{w}}$ by f to give Q_a (111). The mean values of Q_a are then compared with $E(Q)$ as well as with $Q_{\hat{w}}$.

When the weights are estimated from the simulated data, the mean value from the simulations $\overline{Q_{\hat{w}}}$ is larger than the expected value for each of the three choices of w_1 for both values of n (Figures 56–58). This difference is small when n=50, but is much larger when n=5. The increase is only slight when n=50 since f is approximately 1, but much larger when n=5 when f is equal to 2.

The results of the simulations show that $\overline{Q_a}$ provide better approximations to $E(Q)$ than $Q_{\hat{w}}$. Table 47 indicates that the adjustment to Q is particularly good when n = 50, but less so when n=5 where it tends to overcompensate for the large inflation in the statistic. These results, therefore, show that on average the statistic calculated in practice is closer to $fE(Q)$ than it is to $E(Q)$. Hence, the null distribution of $Q_{\hat{w}}$ will not be χ^2_{k-1} and so the test will be incorrect

Following on from the increase in the value of the expectation of Q caused by the estimation of w_i, the power of the test is also increased (Figures 59–61). Again this increase is small when n=50, but large when n=5. It can also be seen that for any given value of n, the absolute increase in power is always greatest when w_1=90, and hence the greatest difference overall occurs when n=5 and w_1=90 (Figure 61). The power of the test decreases as the weights allocated to the studies in the meta-

Figure 56: Mean value of the test statistic Q from the simulations where the weights are estimated from different sample sizes when w_1=10 and $\sum_{i-1}^{k} w_i$=100



Key

– expected value of Q

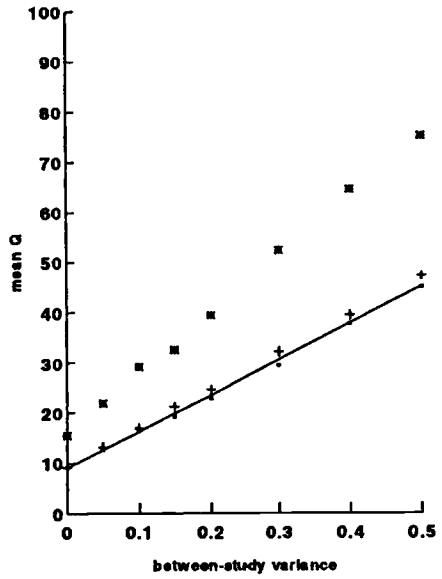
· true weights

+ estimated weights (n=50)

* estimated weights (n=5)

1000 simulations at each point

Figure 57: Mean value of the test statistic Q from the simulations where the weights are estimated from different sample sizes when w_1=50 and $\sum_{i=1}^{k} w_i$=100



Key

– expected value of Q

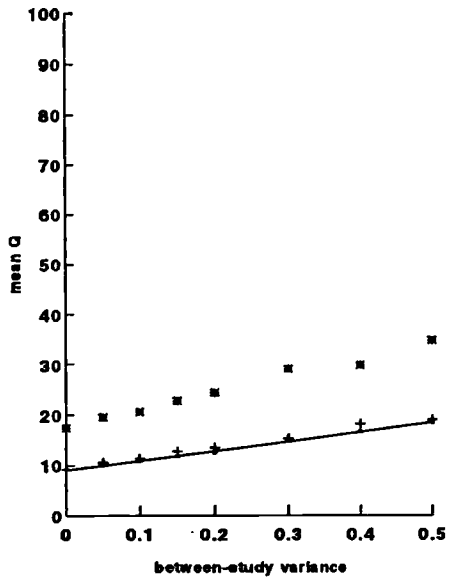
· true weights

+ estimated weights (n_i=50)

* estimated weights (n_i=5)

1000 simulations at each point

Figure 58: Mean value of the test statistic Q from the simulations where the weights are estimated from different sample sizes when w_1=90 and $\sum_{i=1}^{k} w_i$=100



Key

– expected value of Q

· true weights

+ estimated weights (n=50)

* estimated weights (n=5)

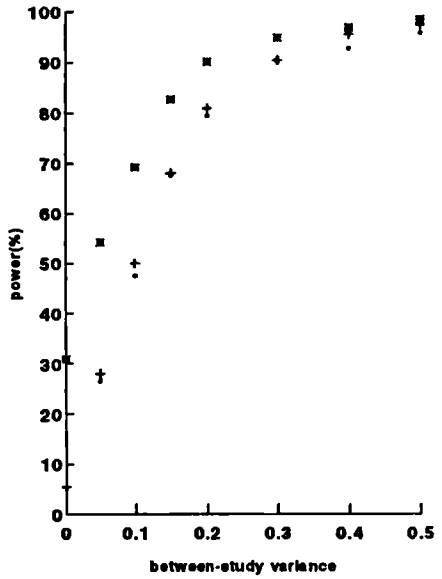
1000 simulations at each point

Table 47: A comparison of the standard test statistic for heterogeneity $Q_{\hat{w}}$ calculated in practice and the adjusted Q_a with $E(Q)$

Weight given to trial 1 w_1	σ_B^2	Number of observations in each trial (n_i)					
		50			5		
		Values from simulations		Analytic	Values from simulations		Analytic
		$\overline{Q_{\hat{w}}}$	$\overline{Q_a}$	$E(Q)$	$\overline{Q_{\hat{w}}}$	$\overline{Q_a}$	$E(Q)$
10	0	9.213	8.837	9.000	15.262	7.631	9.000
	0.05	14.085	13.510	13.500	22.936	11.468	13.500
	0.10	18.648	17.887	18.000	28.501	14.251	18.000
	0.15	23.426	22.470	22.500	39.249	19.625	22.500
	0.20	28.267	27.113	27.000	46.729	23.365	27.000
	0.30	37.890	36.343	36.000	62.162	31.081	36.000
	0.40	46.913	44.999	45.000	78.618	39.309	45.000
	0.50	55.269	53.013	54.000	91.153	45.577	54.000
50	0	9.160	8.786	9.000	15.503	7.751	9.000
	0.05	13.191	12.653	12.611	21.835	10.917	12.611
	0.10	16.856	16.168	16.222	29.172	14.586	16.222
	0.15	21.138	20.275	19.833	32.564	16.282	19.833
	0.20	24.634	23.629	23.444	39.374	19.687	23.444
	0.30	32.155	30.843	30.667	52.414	26.207	30.667
	0.40	39.378	37.771	37.889	64.567	32.283	37.889
	0.50	47.218	45.291	45.111	75.153	37.577	45.111
90	0	9.244	8.867	9.000	17.381	8.691	9.000
	0.05	10.527	10.097	9.944	19.550	9.775	9.944
	0.10	11.370	10.906	10.889	20.600	10.300	10.889
	0.15	12.748	12.228	11.833	22.697	11.349	11.833
	0.20	13.431	12.883	12.778	24.354	12.571	12.778
	0.30	15.299	14.675	14.667	29.052	14.526	14.667
	0.40	18.176	17.434	16.556	29.797	14.899	16.556
	0.50	18.923	18.151	18.444	34.652	17.326	18.444

analysis become more different as was observed when investigating the power with known weights (Chapter 4).

Figure 59: Power of the test statistic Q where the weights are estimated from different sample sizes when w_1=10 and $\sum_{i=1}^{k} w_i$=100



Key

· true weights

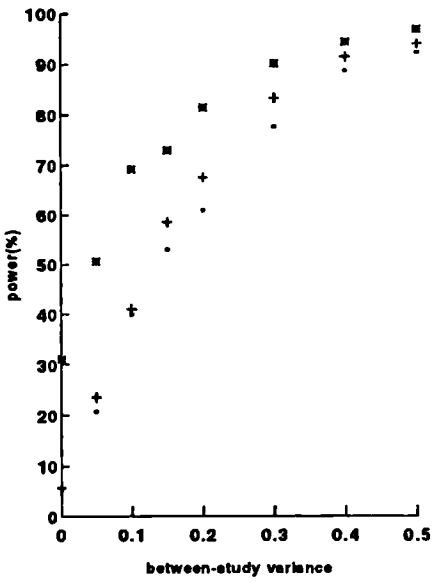
+ estimated weights (n=50)

* estimated weights (n=5)

1000 simulations at each point

When n=5, there is a large increase in the number of tests producing significant results when there is in fact no heterogeneity present, that is an increase in the Type-I error. This value rises from 5%, when the true weights are used, to between 30% and 40%, when the weights are estimated from 5 observations (Figures 59–61). These results indicate that the test for heterogeneity is not valid when n is small. This is supported by the calculation of confidence intervals for the difference in power between the case where the weights are known and the case where weights are estimated

Figure 60: Power of the test statistic Q where the weights are estimated from different sample sizes when w_1=50 and $\sum_{i=1}^{k} w_i$=100



Key

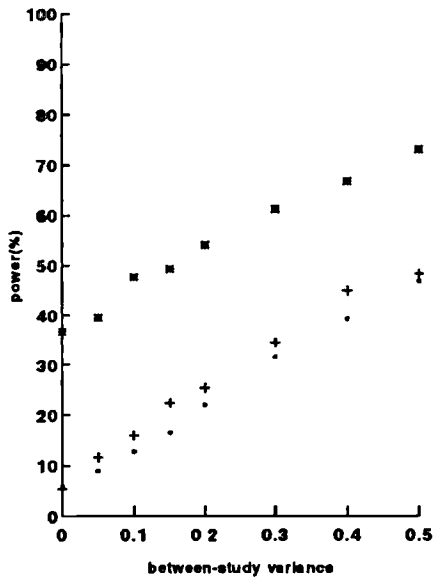
· true weights

\+ estimated weights (n=50)

* estimated weights (n=5)

1000 simulations at each point

Figure 61: Power of the test statistic Q where the weights are estimated from different sample sizes when w_1=90 and $\sum_{i=1}^{k} w_i$=100



Key

. true weights

+ estimated weights (n−50)

* estimated weights (n=5)

1000 simulations at each point

(Table 48). The confidence intervals all indicate an increase in power when n=5. In contrast, this difference in power is much smaller when n=50. Hence, it may be sensible to use Q_a as the test statistic for heterogeneity as this has a null distribution which is at least closer to the χ^2_{k-1} distribution. Although caution is then necessary as the test is even lower in power due to the over correction for the inflation, particularly when n is small.

5.2.3 Between-study variance

This section includes an investigation of the mean of the simulated between-study variances calculated using the D&L method of moments $\overline{\hat{\sigma}^2_{B\hat{w}}}$ and based on the assumption that the weights are known with the true between-study variance. When n_i=n for all i, as in the simulation examples, the adjusted estimate of the between-study variance $\hat{\sigma}^2_{B_a}$ which takes into account the estimation of the weights to some extent at least is given in (120). For each example, the mean simulated value using this alternative estimate $\overline{\hat{\sigma}^2_{B_a}}$ is then also compared with the true σ^2_B and with the standard estimate $\overline{\hat{\sigma}^2_{B\hat{w}}}$ in order to see whether it does in fact offer an improvement.

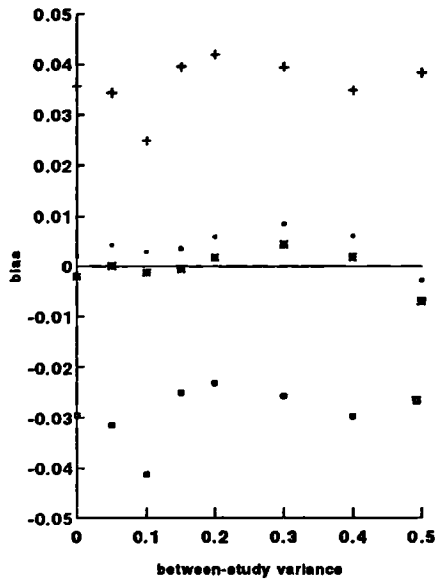
Since $\hat{\sigma}^2_{B\hat{w}}$ is calculated using the value of $Q_{\hat{w}}$ obtained using estimated weights, this causes $\hat{\sigma}^2_{B\hat{w}}$ to be biased. In all six simulated examples and for all σ^2_B, the mean of the simulated $\hat{\sigma}^2_{B\hat{w}}$ was larger than the true value (Figures 62–64). It may be observed from these plots that the bias of the D&L estimator remains approximately constant over all values of σ^2_B. This observed bias (i.e. $(\overline{\hat{\sigma}^2_{B\hat{w}}} - \sigma^2_B)$) was very large, at around 0.2 for the case where w_1=90 and n=5, while in all the other five situations the bias was below 0.05. The adjusted estimate $\hat{\sigma}^2_{B_a}$ does perform better than the D&L estimator in that the average observed bias for each example is smaller (Figures 62–64). There is still some consistent deviation from the true value of σ^2_B when n=5, but the adjusted estimate consistently underestimates, rather than overestimates, the between-study variance. This is due to the overcompensation for the inflation in Q

Table 48: Differences in power (%) between results from simulations where the true weights were used and those where estimated weights were used

Weight given to trial 1 (w_1)	σ_B^2	Number of observations in each trial (n)			
		50		5	
		Difference (d)	95% C.I. of d	Difference (d)	95% C.I. of d
10	0	0.2	(-1.8,2.2)	25.6	(22.4,28.8)
	0.05	1.5	(-2.4,5.4)	27.7	(23.6,31.8)
	0.10	2.5	(-1.9,6.9)	21.7	(17.5,25.9)
	0.15	0.5	(-3.6,4.6)	15.2	(11.5,18.9)
	0.20	1.5	(-2.0,5.0)	10.7	(7.6,13 8)
	0.30	0.2	(-2.4,2.8)	4.6	(2.3,6.9)
	0.40	2.8	(0.8,4.9)	4.1	(2.2,6.0)
	0.50	1.5	(-0.1,3.1)	2.7	(1.2,4.2)
50	0	0.5	(-1.5,2.5)	25.9	(22.7,29.1)
	0.05	2.8	(-0.8,6.4)	29.9	(25.9,33.9)
	0.10	1.1	(-3.2,5.4)	29.2	(25.0,33.4)
	0.15	5.5	(1.2,9.9)	19.9	(15.8,24.0)
	0.20	6.6	(2.4,10.8)	20.5	(16.6,24.4)
	0.30	5.7	(2.2,9.2)	12.6	(9.4,15 8)
	0.40	2.8	(0.2,5.4)	5.7	(3.3,8.1)
	0.50	1.7	(-0.5,3.9)	4.5	(2.6,6 5)
90	0	-0.5	(-2.5,1.5)	30.8	(27.5,34.1)
	0.05	2.6	(-0.1,5.3)	30.5	(27.0,34.0)
	0.10	3.1	(0.0,6.2)	34.9	(31.2,38.6)
	0.15	5.8	(2.3,9.3)	32.8	(29.0,36.7)
	0.20	3.4	(-0.3,7.1)	32.2	(28.2,36.2)
	0.30	2.9	(-1.2,7.0)	29.7	(25.5,33.9)
	0.40	5.7	(1.4,10.0)	27.2	(23.3,31.7)
	0.50	1.6	(-2.8,6.0)	26.3	(22.2,30.4)

noted in the previous section (Section 5.2.2).

Figure 62: A comparison of the bias of the unadjusted estimates and adjusted estimates of the between-study variance σ_B^2 when w_1=10 for $n-50$ and n=5



Key

. unadjusted estimate of the between-study variance $\hat{\sigma}^2_{B\hat{w}}$ (n=50)

* adjusted estimate of the between-study variance $\hat{\sigma}^2_{Ba}$ ($n-50$)

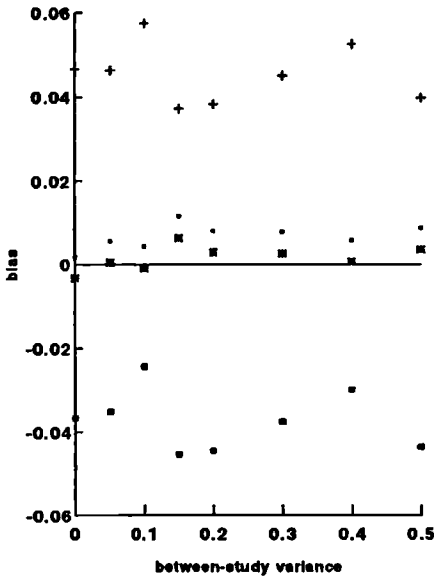
+ unadjusted estimate of the between-study variance $\hat{\sigma}^2_{B\hat{w}}$ (n=5)

o adjusted estimate of the between-study variance $\hat{\sigma}^2_{Ba}$ (n=5)

1000 simulations at each point

Since it is noticeable from the results of the simulations that the bias in the standard estimator is constant for all values of σ_B^2 and an analytical approximation to this bias was found. The D&L estimate of the between-study variance $\hat{\sigma}^2_{B\hat{w}}$, under the assumption that $\hat{w} = fw$, is given by

Figure 63: A comparison of the bias of the unadjusted estimates and adjusted estimates of the between-study variance σ_B^2 when w_1=50 for n=50 and n=5



Key

· unadjusted estimate of the between-study variance $\hat{\sigma}^2_{B\hat{w}}$ (n=50)

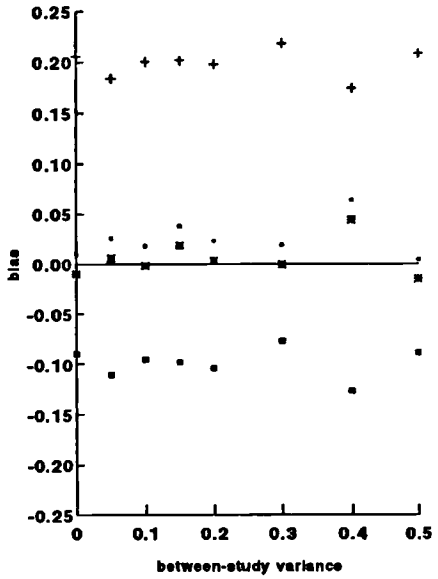
* adjusted estimate of the between-study variance $\hat{\sigma}^2_{B_a}$ (n=50)

+ unadjusted estimate of the between-study variance $\hat{\sigma}^2_{B\hat{w}}$ (n=5)

o adjusted estimate of the between-study variance $\hat{\sigma}^2_{B_a}$ (n=5)

1000 simulations at each point

Figure 64: A comparison of the bias of the unadjusted estimates and adjusted estimates of the between-study variance σ_B^2 when w_1=90 for n=50 and n=5



Key

- unadjusted estimate of the between-study variance $\hat{\sigma}^2_{B\hat{w}}$ (n=50)
* adjusted estimate of the between-study variance $\hat{\sigma}^2_{Ba}$ (n=50)
+ unadjusted estimate of the between-study variance $\hat{\sigma}^2_{B\hat{w}}$ (n=5)
▫ adjusted estimate of the between-study variance $\hat{\sigma}^2_{Ba}$ (n=5)

1000 simulations at each point

$$\hat{\sigma}^2_{Bf} = \frac{fQ-(k-1)}{f\left(\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}\right)} \tag{123}$$

and the true between-study variance assuming that weights are known is σ^2_B. Hence, the approximate bias is

$$\frac{fQ-(k-1)}{f\left(\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}\right)} - \frac{Q-(k-1)}{\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}} = \frac{(f-1)(k-1)}{f\left(\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}\right)} \tag{124}$$

This analytical result shows that the bias is not dependent on the amount of heterogeneity and is therefore constant across all values of σ^2_B (Table 49), as observed.

Table 49: Comparison of the observed bias of the standard D&L estimator and the approximate analytic bias

Number in each group (n)	Weight given to trial 1 (w_1)	Observed bias of $\hat{\sigma}^2_{B\hat{w}}$ ($\overline{\hat{\sigma}^2_{B\hat{w}}} - \sigma^2_B$)	Approximate analytic bias
50	10	0.0038	0.0041
	50	0.0067	0.0051
	90	0.0190	0.0195
5	10	0.0360	0.0500
	50	0.0392	0.0623
	90	0.1989	0.2382

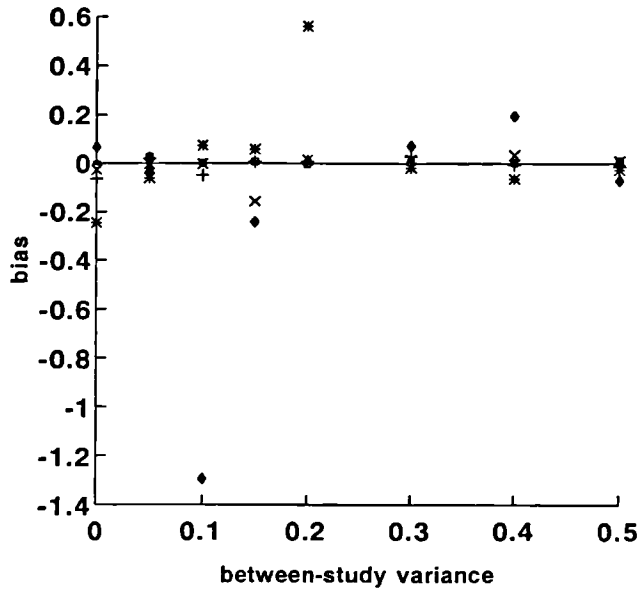
5.2.4 Random effects model

The random effects estimate of treatment effect $\hat{\theta}_{r\hat{w}}$ obtained in practice is considered in order to check for unbiasedness. The standard variance $1/\sum_{i=1}^{k} \hat{w}^*$ of this estimate is then compared with the true sample variance obtained from the simulations $\hat{var}(\hat{\theta}_{r\hat{w}})$. No satisfactory adjusted variance for the random effects estimate was found to compare with these values. The calculation of the variance of a random effects estimate is more complicated than that of the fixed effect estimate because the expectation of a single weight $1/\hat{w}_i^*$ proves difficult to obtain. The estimated weight contains both the biased estimate w_i and that of σ_B^2; these two different effects working on the variance cannot be separated.

As for the fixed effect method, the results from the simulations indicate that estimating the weights causes no systematic bias in the estimate of the overall treatment effect (Figure 65). This is again due to the fact that the random effects estimate is still a weighted average of the θ_i and the changes in the weights do not affect the unbiasedness of the estimate.

The mean simulated variances $\overline{var(\hat{\theta}_{r\hat{w}})}$ using the standard estimate of the variance are, in general, slightly larger than the theoretical variances $1/\sum_{i=1}^{k} w_i^*$ based on the assumption of known weights in the examples where w_1=10 and w_1=50 (Tables 50 and 51). The cases where n=5 produce variances which are on the whole larger than when n=50, with the increase when n=50 being very small indeed. In the example where w_1=90 (Table 52), the estimates are too erratic to enable any firm conclusions to be drawn. The general increase in variance is due to the mean between-study variance estimate being larger than the true value because there is a constant positive bias over all values of σ_B^2. Since the variance is calculated using $1/\sum_{i=1}^{k} 1/(\hat{v}_i + \hat{\sigma}^2_{B\hat{w}})$ and $(\hat{v}_i + \hat{\sigma}^2_{B\hat{w}})$ will be greater than $(v_i + \sigma_B^2)$ due to the additional bias, then $1/(\hat{v}_i + \hat{\sigma}^2_{B\hat{w}})$ will be less than $1/(v_i + \sigma_B^2)$. Hence, the reciprocal of the sum of $1/(\hat{v}_i + \hat{\sigma}^2_{B\hat{w}})$ will be greater than the reciprocal of the sum of $1/(v_i + \sigma_B^2)$.

Figure 65: Plot showing the bias in the random effects estimate of the overall treatment effect $(\overline{\hat{\theta}_{r\hat{w}}} - \theta)$ against the between-study variance



Key

· w_1=10, n_i=50 ▫ w_1=10, n_i=5

\+ w_1=50, n_i=50 × w_1=50, n_i=5

* w_1=90, n_i=50 ♦ w_1=90, n_i=5

1000 simulations at each point

Table 50: Comparison of the standard estimated variance for the random effects estimate of treatment effect with the mean from the simulations and the standard analytic result when $\sum_{i=1}^{k} w_i$=100 and w_1=10

σ_B^2	Number of observations in each trial (n)				Standard analytic variance $(1/\sum_{i=1}^{k} w_i^*)$
	50		5		
	$\overline{(var(\hat{\theta}_{r\hat{w}}))}$	$\hat{var}(\hat{\theta}_{r\hat{w}})$	$\overline{(var(\hat{\theta}_{r\hat{w}}))}$	$\hat{var}(\hat{\theta}_{r\hat{w}})$	
0	0.01160	0.01006	0.01138	0.01255	0.010
0.05	0.01572	0.01638	0.01647	0.01541	0.015
0.10	0.02033	0.02084	0.02073	0.02179	0.020
0.15	0.02525	0.02464	0.02739	0.02801	0.025
0.20	0.03053	0.03079	0.03285	0.03136	0.030
0.30	0.04075	0.03707	0.04281	0.04005	0.040
0.40	0.05052	0.05364	0.05251	0.05030	0.050
0.50	0.05967	0.05758	0.06314	0.06201	0.060

Table 51: Comparison of the standard estimated variance for the random effects estimate of treatment effect with the mean from the simulations and the standard analytic result when $\sum_{i=1}^{k} w_i$=100 and w_1=50

σ_B^2	Number of observations in each trial (n)				Standard analytic variance $(1/\sum_{i=1}^{k} w_i^*)$
	50		5		
	$\overline{(var(\hat{\theta}_{r\dot{w}}))}$	$\hat{var}(\hat{\theta}_{r\dot{w}})$	$\overline{(var(\hat{\theta}_{r\dot{w}}))}$	$\hat{var}(\hat{\theta}_{r\dot{w}})$	
0	0.01350	0.01251	0.01485	0.01704	0.01000
0.05	0.01905	0.02130	0.02039	0.02300	0.01872
0.10	0.02431	0.02755	0.02729	0.02757	0.02471
0.15	0.03047	0.03381	0.03057	0.03467	0.03016
0.20	0.03539	0.03612	0.03640	0.03940	0.03542
0.30	0.04583	0.05468	0.04770	0.04850	0.04571
0.40	0.05594	0.05928	0.05887	0.05590	0.05587
0.50	0.06641	0.06685	0.06790	0.06659	0.06597

Table 52: Comparison of the standard estimated variance for the random effects estimate of treatment effect with the mean from the simulations and the standard analytic result and the when $\sum_{i=1}^{k} w_i$=100 and w_1=90

σ_B^2	Number of observations in each trial (n)				Standard analytic variance $(1/\sum_{i=1}^{k} w_i^*)$
	50		5		
	$\overline{(var(\hat{\theta}_{r\hat{w}}))}$	$\hat{var}(\hat{\theta}_{r\hat{w}})$	$\overline{(var(\hat{\theta}_{r\hat{w}}))}$	$\hat{var}(\hat{\theta}_{r\hat{w}})$	
0	0.03599	0.02872	0.05665	0.05601	0.01000
0.05	0.04722	0.05414	0.06079	0.07336	0.03870
0.10	0.05343	0.07288	0.07022	0.09356	0.05556
0.15	0.06422	0.08625	0.07676	0.09092	0.06767
0.20	0.06931	0.10170	0.08340	0.09623	0.07741
0.30	0.08301	0.11530	0.09858	0.10805	0.09333
0.40	0.10204	0.12545	0.10600	0.12383	0.10689
0.50	0.10778	0.12227	0.12243	0.13701	0.19926

The estimates of the true variance $\hat{var}(\hat{\theta}_{r\hat{w}})$ obtained from the 1000 simulated values of $\hat{\theta}_{r\hat{w}}$ are nearly all larger than $1/\sum_{i=1}^{k} w_i^*$ (Tables 50–52). There are three individual exceptions, however, in the example where n=50 and w_1=10 (Table 50), but these differences are small enough to be regarded as being due to sampling error. Hence, assuming that $1/\sum_{i=1}^{k} w_i^*$ underrepresents the variation in $\hat{\theta}_{r\hat{w}}$ then the results show that the variance obtained in practice $1/\sum_{i=1}^{k} \hat{w}_i^*$ may often be closer to $\hat{var}(\hat{\theta}_{r\hat{w}})$ than to $1/\sum_{i=1}^{k} w_i^*$. However, no more definite conclusions can be drawn since there is no theory to support the findings and furthermore, the results are rather erratic.

5.3 Conclusions

The investigations outlined in this chapter have shown that for qualitative outcome measures the fact that the weights are being estimated can affect the results of the meta-analysis. This is due to some extent to the fact that the expectation of a single estimated weight $\hat{w}_i$ does not equal w_i, but rather $f_i w_i$.

The fixed effect estimate of the overall treatment effect remains unbiased, but the corresponding variance term used in practice, that is $1/\sum_{i=1}^{k} \hat{w}_i$, is too small. This is not a problem when the number of observations in each trial n_i is large since in such circumstances the decrease is negligible. However, it does become an issue when n_i in each trial is very small and the decrease in the variance could cause too definite conclusions to be drawn about the true treatment effect. The adjusted estimate derived in this chapter, $var_a(\hat{\theta}_{f\hat{w}})$, is a better approximation of the true variance, although it performs better in certain cases than in others. However, it does always produce a value which is larger than $1/\sum_{i=1}^{k} w_i$ which is at least an improvement over $1/\sum_{i=1}^{k} \hat{w}_i$. Using the adjusted variance of the fixed effect treatment effect will lead appropriately to a more cautious interpretation of the data.

The test statistic calculated in practice, that is $Q_{\hat{w}}$, is inflated due to the

estimation of the weights. As for the variance of the fixed effect estimate, the effect is very small for large n_i, but is very large when n_i are small. For the case where n_i=n for all i, an adjusted test statistic may be obtained by dividing $Q_{\hat{w}}$ by f. This brings the test statistic towards $E(Q)$. Certainly Q_a is closer to $E(Q)$ than $Q_{\hat{w}}$, although for small n the adjustment overcompensates for the inflation and the test statistic is too small. In practice using $Q_{\hat{w}}$ means that the power of the test is artificially increased; the null distribution of the test statistic is not χ^2_{k-1} and the test is not valid. This is again a particular concern when n_i is small. Furthermore, the inflation increases as w_1 increases. Q_a offers some improvement since it brings the null distribution of the test statistic closer to χ^2_{k-1}, although the possibility of underestimating the extent of the heterogeneity is then a danger when n_i is small.

The results observed for Q then follow through to influence the estimate of the between-study variance. Using the standard D&L moment estimate leads to an overestimation of the between-study variation. This bias is constant over all values of the between-study variance for any given example. However, the extent of the bias is dependent on the allocation of the weight, where the more uneven the allocation of the weight the greater the bias. The adjusted estimate proposed in this chapter $\hat{\sigma}^2_{Ba}$ for the case where n_i=n for all i was again an improvement over the standard estimate, although not ideal. The reduction in Q_a when n is small leads to the underestimation of σ^2_B.

The conclusions that can be drawn for the random effects model are rather limited. The estimate of the treatment effect appears to remain unbiased, although the variance clearly is larger than $1/\sum_{i=1}^{k} w_i^*$ due to the estimation of both v_i and σ^2_B. The variance calculated in practice, that is $1/\sum_{i=1}^{k} \hat{w}_i^*$ is also larger than $1/\sum_{i=1}^{k} w_i^*$ due to the influence of the inflated estimate of σ^2_B. However, there is no theory to state that $1/\sum_{i=1}^{k} \hat{w}_i^*$ is a reasonable estimate of the true variance.

Hence, overall, in most practical situations where n_i is large the standard es-

timates perform adequately. However, in such cases, for example where n=50, the adjusted estimates do perform even better, although the adjustments are so small that they are most unlikely to make any real difference to the conclusions drawn. The case where n_i is small is where problems occur and adjustments are more important. Unfortunately, when n was equal to 5 the simulated results indicated that the adjustments were less good. However, improvements were still seen with regards to the $var(\hat{\theta}_f)$, Q and $\hat{\sigma}_B^2$.

As always with a simulation study, the results are not necessarily generalisable to situations which were not investigated. Further simulation examples were carried out in cases where n_i was allowed to vary within a meta-analysis. These showed that where only some of the trials have small numbers of observations and others have large numbers problems can still occur. Hence, if a meta-analysis includes just one small study an impact on the results is possible. The results do not apply to binary outcomes and hence, there is scope for this work to be extended, especially as it has been shown that the estimation of the weights can affect the meta-analysis results. However, this problem may, perhaps, be more usefully approached by consideration of the full likelihood method of van Houwelingen et al. [45] (Section 2.4) which allows for the estimation of the weights for binomial outcome measures. A comparison, using simulated data, of the standard results with those obtained from the full likelihood method when numbers of observations in each trial is small would be informative.

6 Analysis of Data From the British Family Heart Study

In this chapter the application of meta-analysis techniques to the analysis of a paired cluster randomised trial is described. Data from the British family heart study [112] is then analysed using such methods. Section 6.1 describes the design and conduct of the British family heart study, while Section 6.2 illustrates how the random effects meta-analysis methods can be applied to such studies. Section 6.3 then presents the results for a selection of outcome measures and provides an in-depth consideration and investigation of heterogeneity for one continuous outcome measure (level of cholesterol) and one binary outcome measure (prevalence of smoking). Section 6.4 contains a discussion of the problems of analysing a multicentre trial and trying to account for heterogeneity. Furthermore, due to the multiple endpoints recorded in the British family heart study, Section 6.5 considers the problem of multiple testing and briefly introduces the concept of a multivariate meta-analysis.

6.1 Introduction to the British Family Heart Study

The aim of the British family heart study was to measure the change in cardiovascular risk factors achievable in families over one year by the implementation of a cardiovascular screening and lifestyle intervention programme based in general practice [112]. The intervention programme was nurse led, with a different research nurse being allocated to each intervention practice. Research nurses were recruited locally and trained centrally before commencing the study.

Two general practices in each of 13 towns in Britain (10 in England, 1 in Scotland and 2 in Wales) with a population of between 50 000 and 100 000 at the 1981 census were identified. The pair of practices within each town were matched so

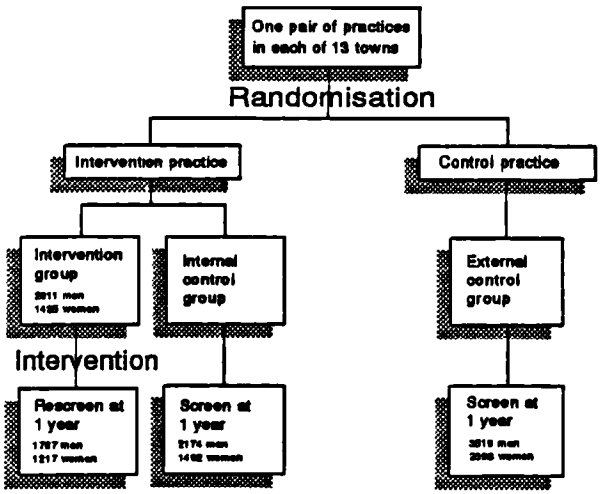
that they had similar sociodemographic characteristics. The two practices within each town were then randomly allocated to be either the intervention practice or the control practice (Figure 66). Families recruited to take part in the study were identified through the male partner who, in order to be eligible, had to be aged between 40 and 59 years. In each intervention practice and each control practice, all men aged 40–59 years were randomly ordered within five year age bands. Furthermore, in the intervention practices, each age band was randomly split into two equal sized groups, one of which became the intervention group and the other the internal control group (Figure 66). The families randomised to the intervention group were then contacted by the practice research nurse in the order given by the five year age band lists. Contacts were made at the same rate within each five year band. The families were screened and subsequently offered lifestyle intervention and follow up.

All family members attending the initial visit were screened, but only men and their partners were followed up. During the initial screening interview, demographic, lifestyle, and medical information were collected. Measurements of height, weight, body mass index (weight/height^2), carbon monoxide concentration in the breath, blood pressure, and random blood concentration of total cholesterol and glucose in a finger prick sample were also obtained.

A coronary risk score was then derived and participants were told in which decile of the distribution of risk for coronary heart disease they were relative to other men (or women) of the same age. The risk score was recorded and relevant lifestyle changes were individually negotiated with the research nurse relating to smoking, weight, healthy diet, alcohol consumption and exercise. The frequency of the follow-up visits was determined by the overall coronary risk score and specific individual factors. The greater the risk, then the more frequent the visits.

Rescreening of men and their partners in the intervention group then took place one year after the initial screening. Identified families in both the external

Figure 66: Design of the British family heart study showing the numbers of men and women randomised and screened



control practices and the internal control groups were unaware that they were participating in a trial until they were called for screening at the end of the one year intervention period. Hence their first screening coincided with the rescreening of the intervention group. In total, 7 460 men and 5 012 women were included in the study (Figure 66). The smaller numbers in the intervention and internal control groups are due to the population in the intervention general practice being split between these two groups.

6.2 Statistical Methods

The comparison of the intervention group with the internal control group in the British family heart study is of a typical multicentre trial design, in which members of a single population (patients at a single practice) are individually randomised to one of two groups. Hence, meta-analysis methodology applies to this comparison in the same way that it does to the MRC mild hypertension trial (Section 1.3.2). In such situations there may be variability in the effect of the intervention between towns, which may be due to the varying ability and effectiveness of each nurse, the differing general practices and the differences in the patient populations and their attitudes towards changing their lifestyle.

However, the comparison of the intervention group with the external control group is of a paired cluster randomised design. The two general practices within each town are randomised to be either the intervention or the control. Hence, within each town (strata) there are two separate clusters of patients who are to be compared, one being the intervention group and the other being the external control group. It is now shown how this sort of design can also be analysed using meta-analysis methods. The treatment effect in the i^{th} town (i=1,...,k) is denoted by $\hat{\theta}_i$, which may be, for example, a log odds ratio for a binary outcome measure or a difference in means for a continuous outcome.

For a continuous outcome, $\hat{\theta}_i = \overline{y}_{i1.} - \overline{y}_{i2.}$ where $\overline{y}_{ij.}$ is the mean of the n_i individual observations in strata i and treatment group j (j=1 for treatment and j=2 for control). Then the variance of the mean is $var(\overline{y}_{ij.}) = (\sigma_i^2/n_i) + \sigma_B^2$, where σ_i^2 is the within-cluster (within-practice in the case of the British family heart study) variance and σ_B^2 is the between-cluster variance. Hence, the variance of this difference is given by

$$var(\hat{\theta}_i) = var(\overline{y}_{i1.}) - var(\overline{y}_{i2.}) = 2\sigma_B^2 + \left(\frac{\sigma_{i1}^2}{n_{i1}} + \frac{\sigma_{i2}^2}{n_{i2}} \right) \tag{125}$$

and so $2\sigma_B^2$ can be considered as the between-stratum variance in a random effects model, or a between-town variance with respect to the British family heart study. This between-stratum variance may then be estimated using either the D&L moment estimator (Section 1.7.1) or by likelihood methods (Section 2.2), assuming the model

$$\hat{\theta}_i \sim N(\theta_i, (\sigma_{i1}^2/n_{i1}) + (\sigma_{i2}^2/n_{i2})) \tag{126}$$

$$\theta_i \sim N(\theta, 2\sigma_B^2) \tag{127}$$

Hence, by taking a weighted average of the individual within-stratum estimates of treatment effect, where the weights are equal to $1/var(\hat{\theta}_i)$ and $var(\hat{\theta}_i)$ is given by (125), the equivalent of a random effects meta-analysis is obtained (Section 1.7.1).

For a binary outcome, the log odds ratio may be used as a measure of treatment effect $\hat{\theta}_i$ and so the variance of $\hat{\theta}_i$ may easily be obtained where the within-stratum component is given in Section 1.5.1. Alternatively, using the difference in prevalence rates, that is $\hat{\theta}_i = \hat{P}_{i1} - \hat{P}_{i2}$ where $\hat{P}_{i1} = a_i/n_{i1}$ and $\hat{P}_{i2} = c_i/n_{i2}$ and a_i and c_i are the number of positive responses (or events) in treatment groups 1 and 2 respectively (Table 2), means that the within-stratum variance is given by $(a_i b_i/n_{i1}^3) + (c_i d_i/n_{i2}^3)$.

For either outcome measure the meta-analysis methods again follow through with the estimate of the between-stratum variation being obtained either using the D&L moment estimator or using likelihood methods.

In this cluster randomised design there are, in addition to the previously mentioned differences between towns, differences between the two general practice populations within each town. Hence, a greater amount of heterogeneity would be expected in the results for the external control group comparison than for the internal control group comparison and so the internal control group comparison would be expected to produce the more precise results. However, due to the possibility of a transfer of the effect of the lifestyle advice from the intervention to the control group within the practice, the magnitude of the intervention effect may be diluted. Thus, the main statistical comparison laid down in the protocol of the study was that of the intervention group with the external control. Both control group comparisons are considered here and the results compared.

6.3 Results

6.3.1 Overall results

Four separate analyses were carried out in order to see if any further insight into the results or the heterogeneity between towns could be achieved from doing 'parallel' analyses on the same study. The four analyses came about by considering men and women separately and also by looking at comparisons of the intervention group with both the internal and external control groups.

The differences between the intervention and control groups at the one year screening for five cardiovascular risk factors are presented in Table 53. The crude summary measure (prevalence or mean) for each group is given, together with the

estimate of the overall difference (difference in prevalence or difference in means) between the intervention and the control groups obtained using the standard random effects meta-analysis methods (Section 1.7.1) and the standard error of this difference. It may be seen that both control groups give similar results and these indicate that, in relation to the five risk factors considered, the intervention group were at less risk than either of the control groups. For both men and women smoking prevalence was lower in the intervention group than in either control group. For men the prevalence was about 4% lower in the intervention group than in either of the control groups and these differences were significant at the 5% level. For women the difference was smaller (3% and 3.6%) and less conclusive owing to the larger standard errors associated with the differences.

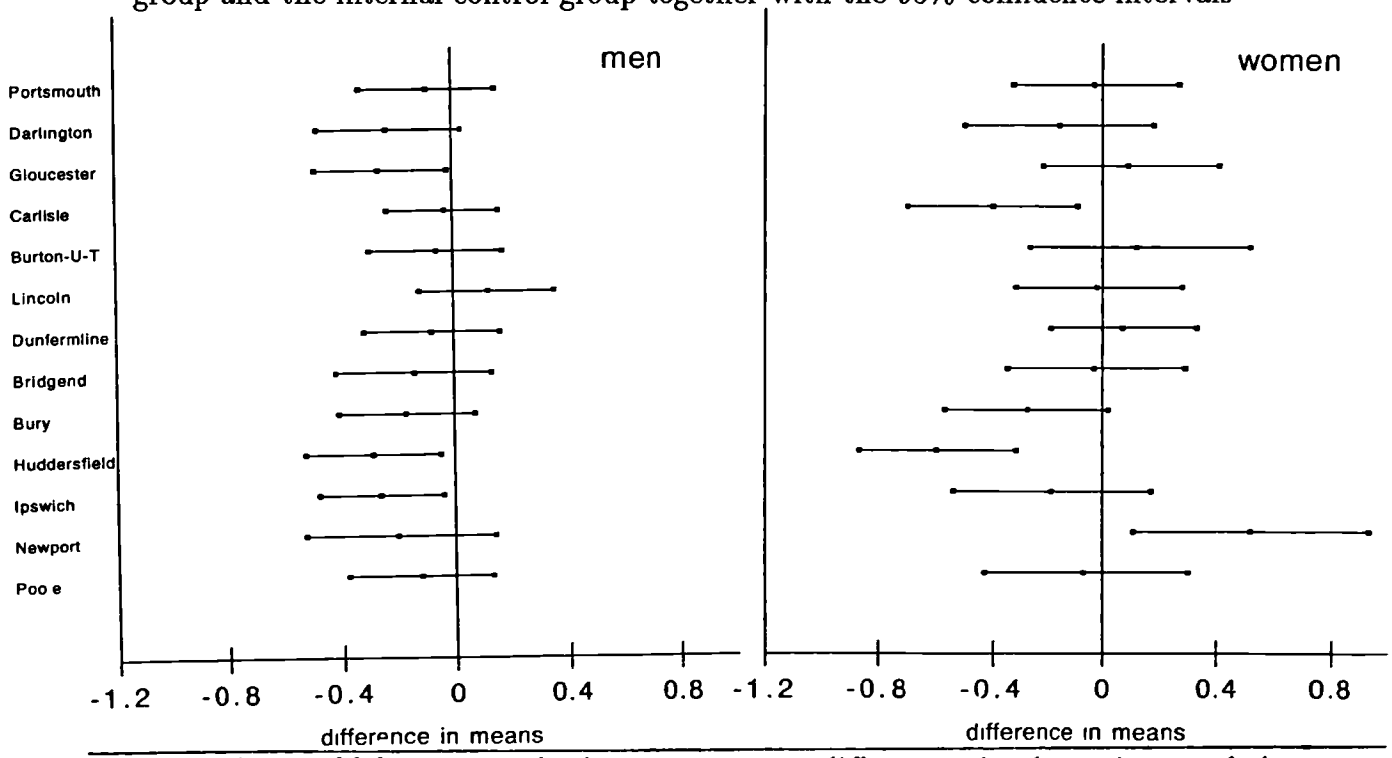
The mean cholesterol level was approximately 0.1mmol/l on average lower in the intervention group than in either control group for both men and women. However, the standard error associated with this difference for women was large enough for the possibility to exist of there being no intervention effect. For both systolic and diastolic blood pressure the means were lower in the intervention group than for either control group for both men and women. This difference was on average around 7mmHg for systolic and 3mmHg for diastolic. The mean weight in the intervention group was also lower, by about 1kg on average for both sexes, than that in either control group.

The fact that all outcomes for all comparisons are in the same direction, tends to add support to the existence of a real intervention effect, even though some differences may not be very large or conclusive. However, the possibility of biases having occurred and influenced the findings should be considered and was in fact discussed in the paper presenting the principal results of the study [112]. For example, the low smoking prevalence observed in the intervention group at the end of the intervention period could have been biased by non-returners or by the under-reporting of current

cigarette smoking in those who did return for rescreening. It was found [112] that the non-returners had a higher smoking rate at baseline than the returners and hence this would have exaggerated the intervention effect with regards to smoking prevalence.

The meta-analysis diagrams for both cholesterol level and smoking prevalence indicate that there is little variation in the estimates of intervention effect across towns for the internal control group comparison for both men and women (Figures 67 and 68). Estimates with approximately equal precisions, as observed in this study, are more likely to occur in a multicentre trial than in a meta-analysis, since as well as the same protocol being followed in each centre, the numbers of patients recruited in each centre will often be roughly comparable. In a meta-analysis, however, the trials may vary greatly with respect to protocol and sample size and thus precision.

Figure 67: Differences in mean cholesterol level (mmol/l) between the intervention group and the internal control group together with the 95% confidence intervals

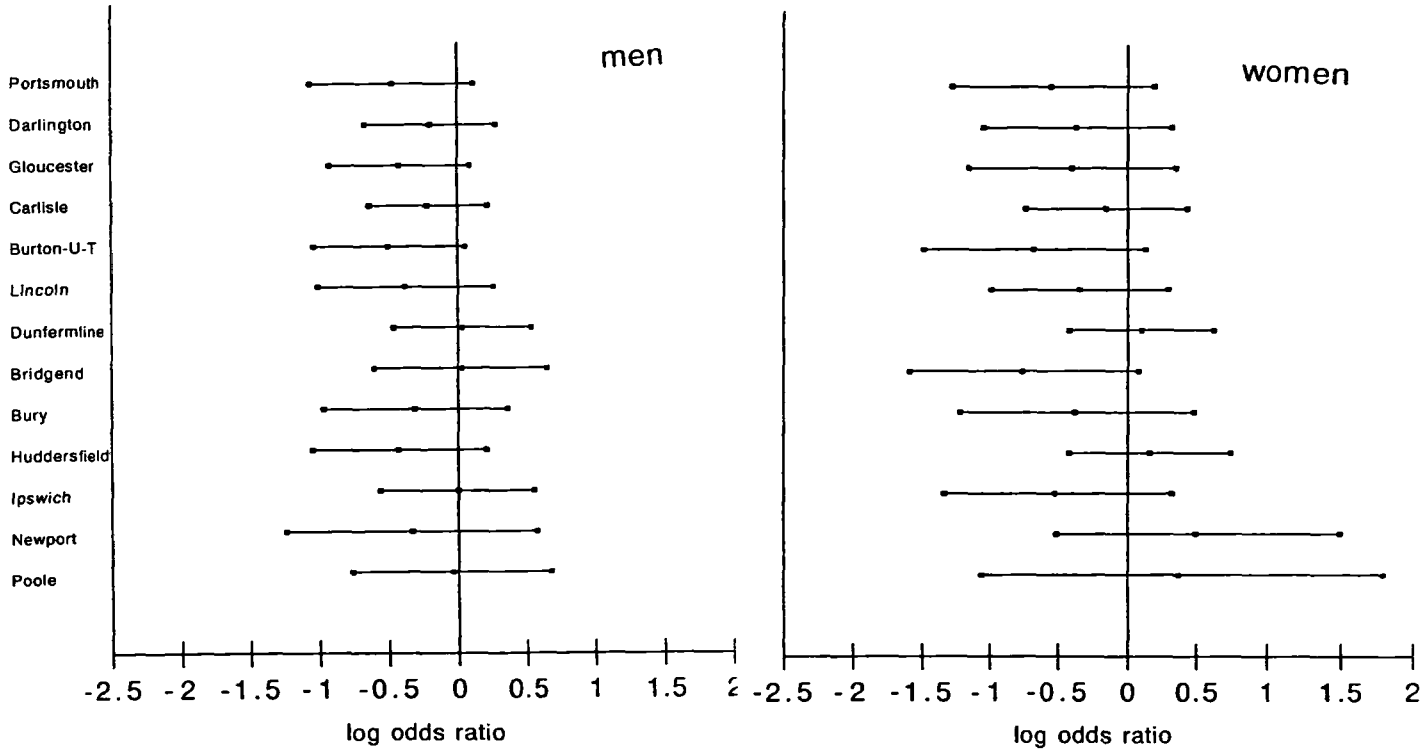


As would be expected, there are greater differences in the estimate of the intervention effect for both cholesterol level and smoking prevalence in both men and

Table 53: Results for five cardiovascular risk factors for the British family heart study

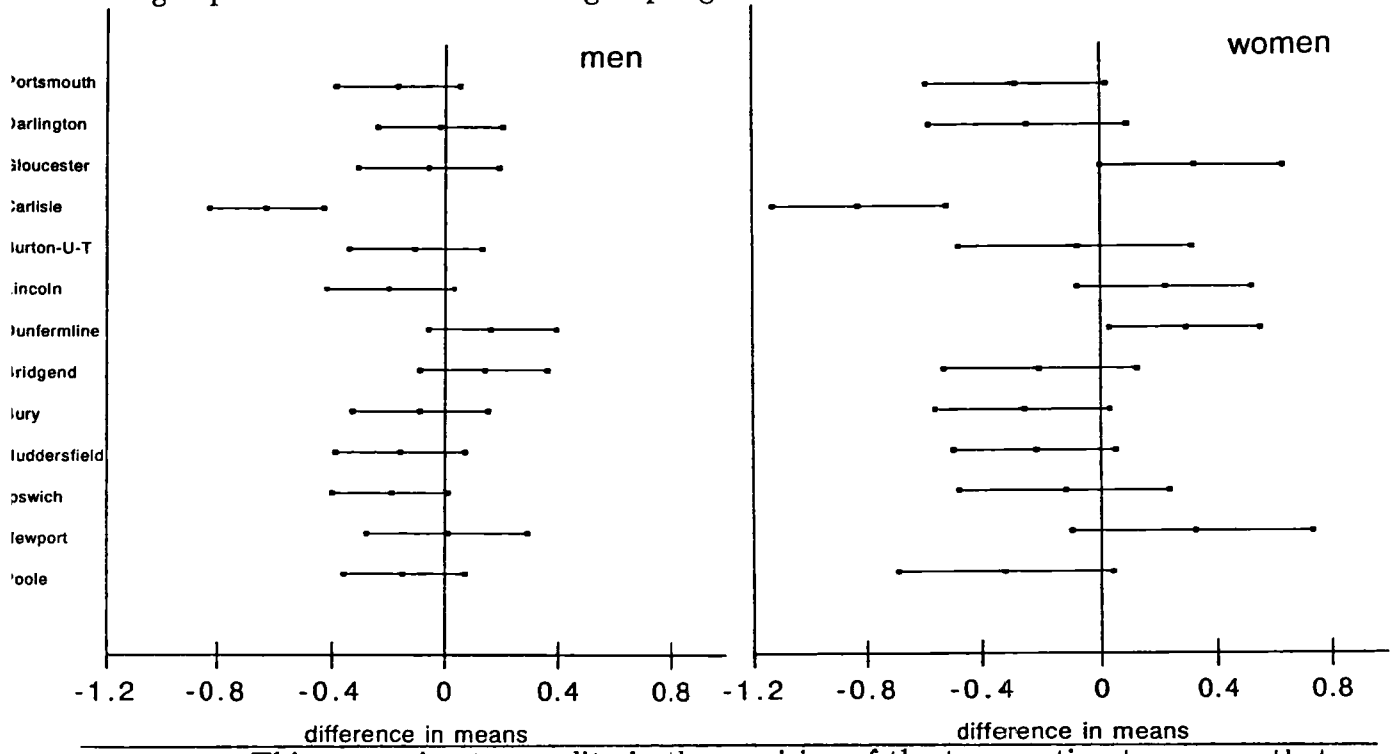
Group	Men		Women	
	crude value	pooled difference (SE)	crude value	pooled difference (SE)
Smoking prevalence (% of subjects)				
Intervention	19.1		17.7	
External control	22.8	-4.1(1.8)	21.2	-3.60(2.1)
Internal control	23.0	-4.1(1.3)	21.5	-3.00(1.5)
Mean blood cholesterol (mmol/l)				
Intervention	5.58		5.48	
External control	5.69	-0.12(0.06)	5.61	-0.12(0.09)
Internal control	5.72	-0.13(0.03)	5.60	-0.09(0.07)
Mean systolic blood pressure (mm Hg)				
Intervention	131.6		123.2	
External control	138.8	-7.5(1.2)	130.8	-7.7(1.4)
Internal control	139.0	-7.3(0.8)	129.6	-6.2(0.9)
Mean diastolic blood pressure (mm Hg)				
Intervention	83.3		78.6	
External control	85.5	-2.5(1.0)	80.7	-2.5(0.9)
Internal control	86.6	-3.5(0.4)	81.3	-3.0(0.4)
Mean weight (kg)				
Intervention	79.55		66.06	
External control	80.70	-1.17(0.36)	66.83	-1.09(0.42)
Internal control	80.76	-1.18(0.43)	66.73	-0.74(0.54)

Figure 68: Differences in prevalence of cigarette smoking between the intervention group and the internal control group expressed as a log odds ratio together with the 95% confidence intervals



women for the external control group comparison (Figures 69 and 70). The precision on each estimate is also more variable, illustrated by the greater variation in the width of the confidence intervals (Figures 69 and 70). However, compared to a meta-analysis, for example the diuretics trials (Figure 1), the variation is still relatively small.

Figure 69: Differences in mean cholesterol level (mmol/l) between the intervention group and the external control group together with the 95% confidence intervals



This approximate equality in the precision of the town estimates means that the weight allocated to each town will be fairly even. For the examples considered here, the percentage weights for the random effects model indicate an even spread of weight across towns (Figures 71 and 72). For men, Carlisle (town 4) is the most informative town for both outcomes and for both control group comparisons, while Dunfermline (town 7) is the most informative for women. There is no single town which dominates the overall estimate of the intervention effect, although a few towns, for example Poole (town 12) and Newport (town 13), contribute less information, particularly to the fixed effect estimates for the internal control group comparison.

Figure 70: Differences in prevalence of cigarette smoking between the intervention group and the external control group expressed as a log odds ratio together with the 95% confidence intervals

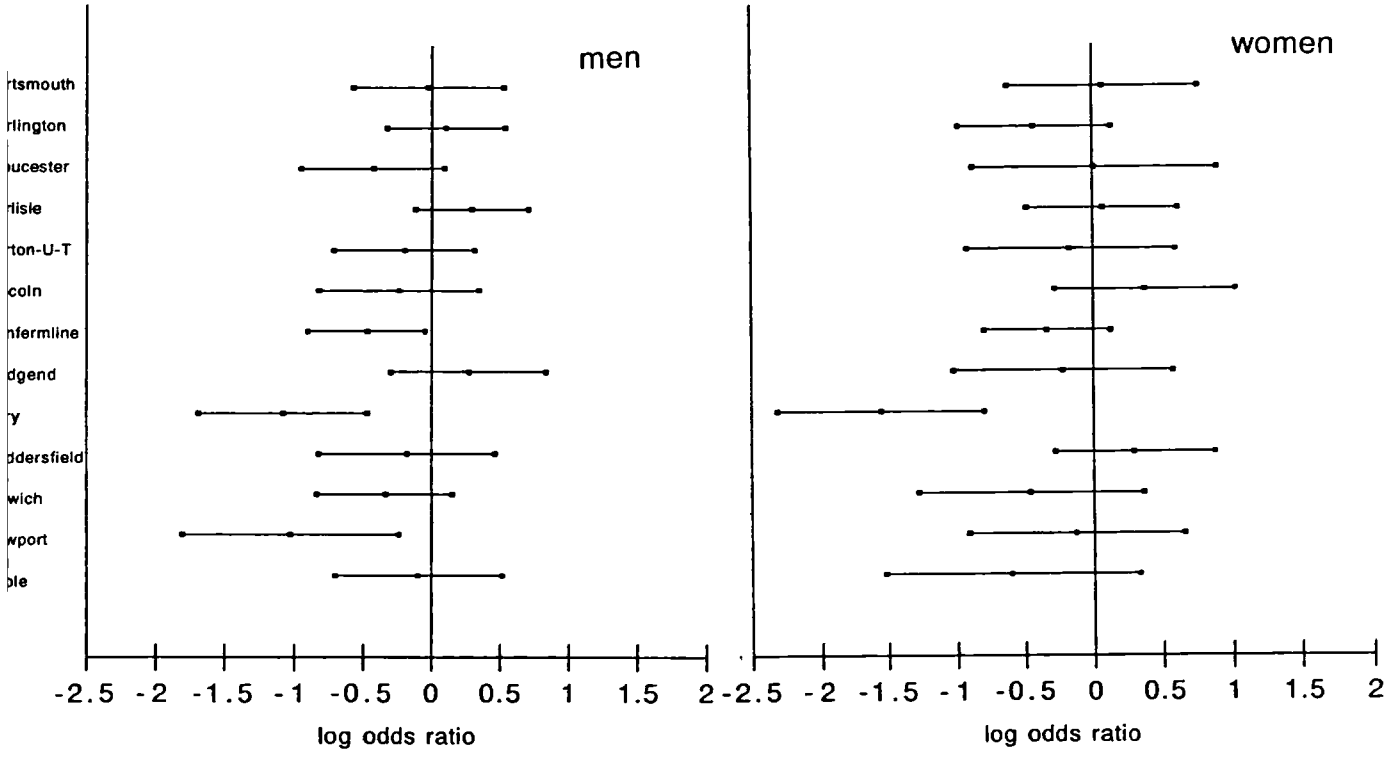
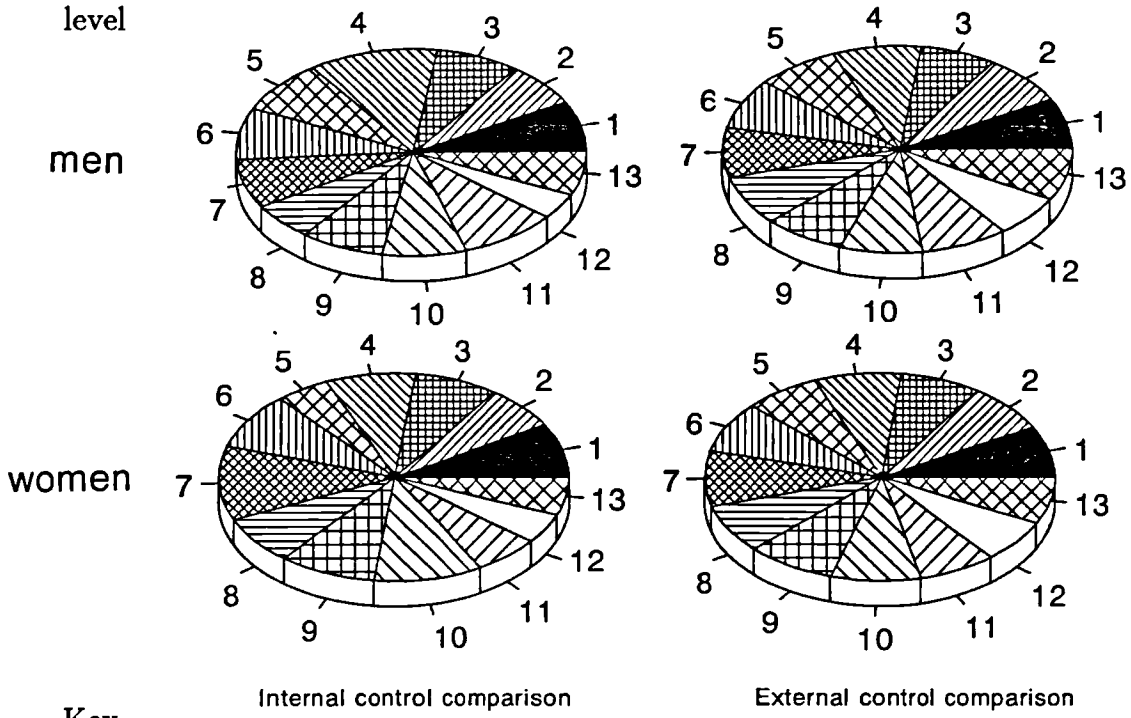


Figure 71: Pie charts showing the percentage weight allocated to each town in the random effects estimate of overall treatment effect for the difference in cholesterol level

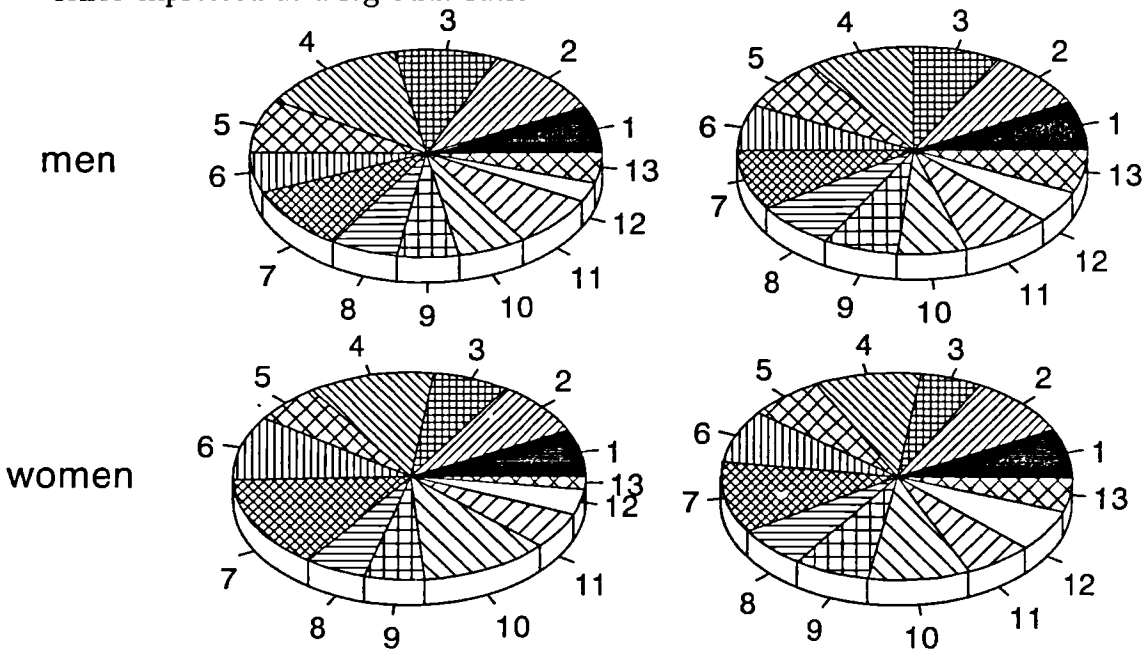


Key

1 Portsmouth, 2 Darlington, 3 Gloucester, 4 Carlisle, 5 Burton-Upon-Trent, 6 Lincoln 7 Dunfermline, 8 Bridgend, 9 Bury, 10 Huddersfield, 11 Ipswich, 12 Newport, 13 Poole

NB. For both internal control group comparison $\hat{\sigma}_B^2=0$

Figure 72: Pie charts showing the percentage weight allocated to each town in the random effects estimate of overall treatment effect for the difference in smoking prevalence expressed as a log odds ratio



Key Internal control comparison External control comparison

1 Portsmouth, 2 Darlington, 3 Gloucester, 4 Carlisle, 5 Burton-Upon-Trent, 6 Lincoln 7 Dunfermline, 8 Bridgend, 9 Bury, 10 Huddersfield, 11 Ipswich, 12 Newport, 13 Poole NB. For both internal control group comparison $\hat{\sigma}_B^2$=0

The mean level of blood cholesterol concentration and the prevalence of current cigarette smoking are now considered in greater detail in Sections 6.3.2 and 6.3.3 respectively. The two aims of these sections are to compare the results obtained using three different meta-analyses, that is the standard inverse-variance fixed effect method (Section 1.5.1), the DerSimonian and Laird random effects method (Section 1.7.1) and the maximum likelihood approach with profile likelihoods used to obtain confidence intervals (Section 2.2), and to investigate heterogeneity from a more practical perspective.

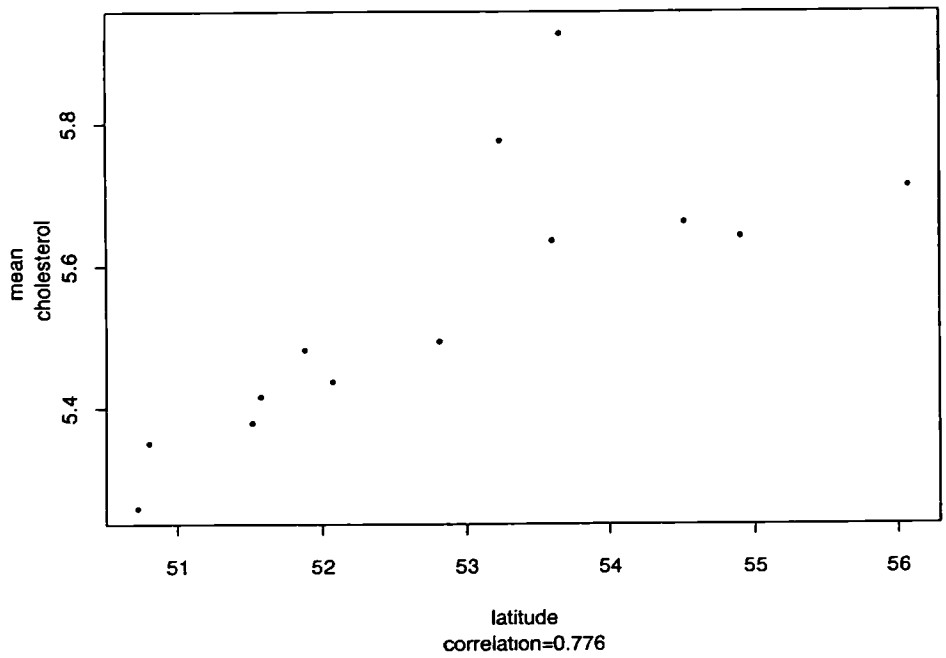
6.3.2 Analysis of cholesterol level

This section focuses on a continuous outcome measure from the family heart study, that is blood cholesterol concentration measured in mmol/l. Differences at the end of the intervention period in the cholesterol level between participants in the intervention group and those in the control group, $\hat{\theta}_i = \overline{y}_{i1.} - \overline{y}_{i2.}$ (1=intervention, 2=control), were considered as the measure of outcome for the analysis. A meta-analysis is carried out followed by a discussion of heterogeneity.

Geographical variation in the level of cholesterol concentration may be expected across towns in this study, since rates of heart disease, and therefore cardiovascular risk factors, are known to vary from region to region. It is also possible that the effect of the lifestyle intervention programme in the British family heart study may vary with baseline cholesterol level and hence with geographical location. Mean cholesterol levels for each group as well as for the differences in mean cholesterol levels were, therefore, plotted against latitude. The results obtained from this study do, at least to some extent, provide evidence of a gradient of cholesterol levels from north to south. For women in the internal control groups there is a particularly strong relationship between mean cholesterol level and geographical location (Figure 73), with a clear increase in mean cholesterol as towns become more northerly. However,

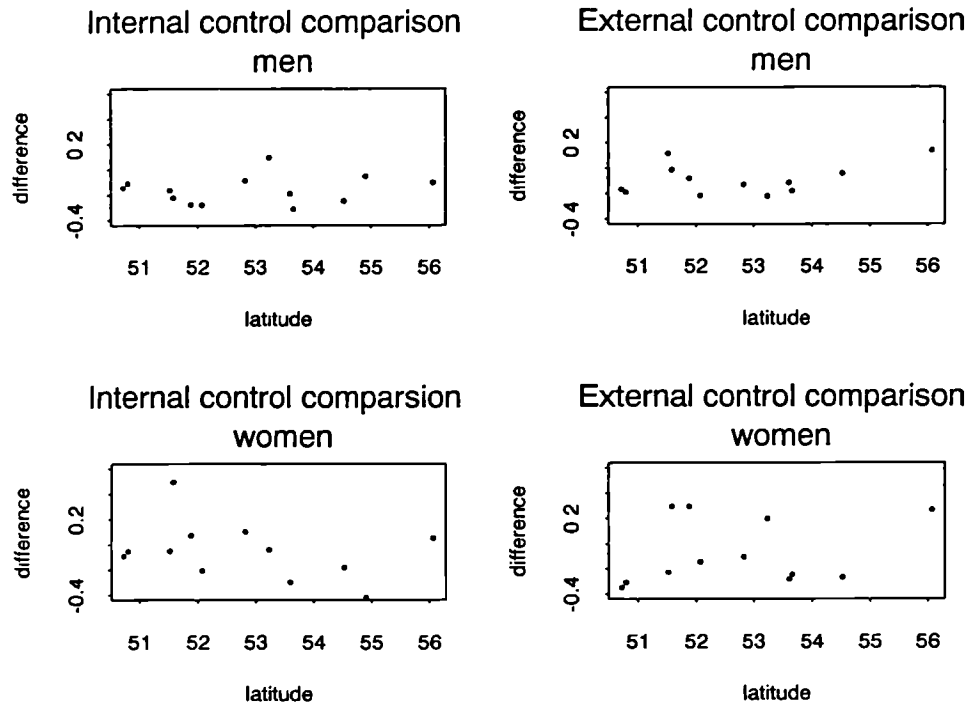
there does not appear to be any relationship between treatment effect and latitude in any of the four comparisons of interest (Figure 74). Hence, it is perhaps reasonable to assume that any regional differences that do exist between centres have been cancelled out when a difference between intervention and control groups is taken. This implies that there is no clear variation according to latitude in the way that different populations respond to intervention.

Figure 73: Mean cholesterol levels in each town for women in the internal control groups plotted against the latitude of the town



Although the concerns regarding the low power of the test for heterogeneity (Chapter 4) and the bias in the estimates caused by the estimation of the weights (Chapter 5) must be considered as potential problems, neither are likely to be important in the analysis of the British family heart study. This is because the numbers of observations in each town is large and the numbers are approximately equal. Hence, the power of the test for heterogeneity will not be particularly low and the estimated weights will only be slightly greater than the true weights on average. When comparing the internal control group and the intervention group, there is no heterogeneity

Figure 74: Differences in mean cholesterol levels in each town plotted against the latitude of the town for all four comparisons



present between towns for the men as measured by Q=10.6 on 12 degrees of freedom. Furthermore, the likelihood ratio test of $\sigma_B^2 = 0$ also indicates a lack of heterogeneity and hence, all three methods produce the same estimate of intervention effect of −0.133mmol/l, indicating a lower mean cholesterol level in the intervention group (Table 54). The 95% confidence interval indicates that this difference is significant at the 5% level. When comparing the intervention group with the external control group, heterogeneity was found to be present (Table 54). Evidence of a significant difference in cholesterol levels (5% level) was detected by all three meta-analysis methods, with the intervention group again having the lower levels. The estimates from the two random effects models were slightly less negative (i.e. smaller difference between groups) than the estimate from the fixed effect model. The confidence intervals were also, of course, wider for the random effects estimates with the widest interval being that calculated from the profile likelihood, whose upper bound was only marginally less than zero (Table 54).

Table 54: Comparison of the results from three different meta-analysis methods for differences in mean blood cholesterol concentration between intervention and control groups in men in the British family heart study

Comparison group	Method	Estimated between-study variance ($\hat{\sigma}_B^2$)	95% C.I. for σ_B^2	Estimated overall effect ($\hat{\theta}$)	95% C.I. for θ
Internal	Fixed	-	-	-0.133	(-0.200,-0.066)
	Random	0.000	-	-0.133	(-0.200,-0.066)
	Likelihood	0.000	(0.000,0.0405)	-0.133	(-0.200,-0.066)
External	Fixed	-	-	-0.126	(-0.190,-0.064)
	Random	0.029	-	-0.117	(-0.229,-0.004)
	Likelihood	0.025	(0.006,0.077)	-0.117	(-0.231,-0.001)

Table 55: Comparison of the results from three different meta-analysis methods for differences in mean blood cholesterol concentration between intervention and control groups in women in the British family heart study

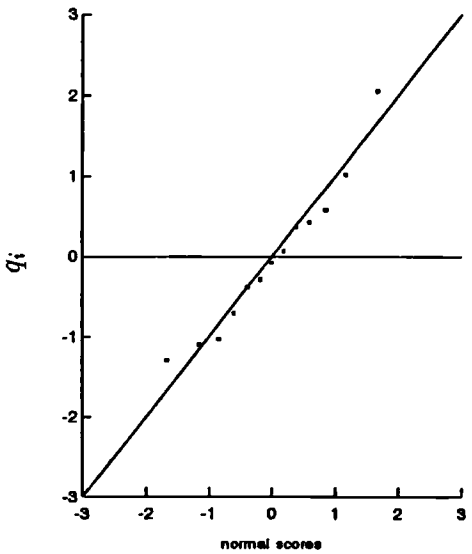
Comparison group	Method	Estimated between-study variance ($\hat{\sigma}_B^2$)	95% C.I. for σ_B^2	Estimated overall effect ($\hat{\theta}$)	95% C.I. for θ
Internal	Fixed	-	-	-0.105	(-0.193,-0.017)
	Random	0.043	-	-0.087	(-0.232,0.057)
	Likelihood	0.037	(0.006,0.130)	-0.087	(-0.299,0.064)
External	Fixed	-	-	-0.111	(-0.199,-0.023)
	Random	0.087	-	-0.113	(-0.297,0.071)
	Likelihood	0.075	(0.026,0.217)	-0.113	(-0.300,0.076)

The results for women (Table 55) are less convincing in favour of the intervention. Although the estimated overall intervention effects were similar for both men and women, the associated confidence intervals were much wider for women than for men. For both comparisons (internal and external control groups) for women the fixed effect analysis provides evidence of a difference in cholesterol levels, but once heterogeneity is taken into account and the confidence intervals are widened, the results are compatible with the possibility of there being no intervention effect. There is greater heterogeneity in the results between towns for women than for men, as well as smaller numbers leading to the wider confidence intervals and less conclusive results.

The q-q plots, both fixed effect (Section 3.1.1) and random effects (Section 3.1.2), were then obtained in order to investigate the form and cause of the heterogeneity observed in three out of the four examples. They were also used to check the validity of the modelling assumptions of normality in conjunction with version (3) of the Anderson-Darling test (Section 3.2.2). However, the test proved of limited use in the examples presented here, since it did not appear to have the power to detect deviations from the model, presumably due to the relatively small number of points. It may be seen from the fixed effect q-q plot (Figure 75) and from the result of the Anderson-Darling test (A^2=0.239, p >0.15) that the intervention effect estimates in each town for the comparison of the intervention group with the internal control group for men are consistent with a normal distribution, that is $\hat{\theta}_i \sim N(\theta, v_i)$. There is clearly no significant heterogeneity in this example and the distributional assumptions of the normally distributed fixed effect model appear to be adequate. However, the corresponding plot (Figure 76) for the external control comparison shows that there is one clear outlying town in these data, rather than the heterogeneity following a random effects model. However, the Anderson-Darling test does not pick this up and produces a nonsignificant result (A^2=1.56, p >0.15). Since the variances, and hence the weights, of each town estimate of intervention effect are fairly equal, then

data following a random effects model plotted using fixed effect q_i would produce an approximate straight line with a gradient steeper than one. It would not produce the type of plot seen in this example, although again the Anderson-Darling test produces a nonsignificant result. Hence, in this particular case, more information is picked up from the plot than by the results of the Anderson-Darling test. By considering the component q_i^2 of the heterogeneity test statistic Q contributed by Carlisle (town number 4), it may be seen that the heterogeneity is due entirely to this single observation. The contribution of this town to Q is $(-4.824)^2$=23.27 and Q calculated without the observation for Carlisle produces a p-value greater than 0.1 when compared to the χ^2_{11} distribution.

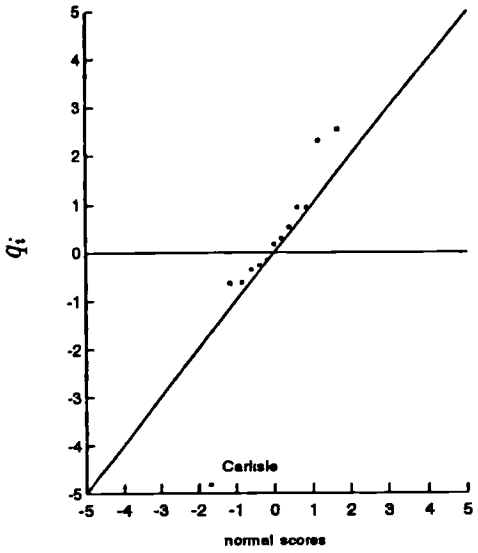
Figure 75: Fixed effect normal plot of q_i for the internal control group comparison of cholesterol levels for men compared with the $N(0,1)$ line



correlation=0.978

The group mean cholesterol levels for Carlisle (Table 56) show why the results for this particular town deviate from the rest and why it produces an outlying observation in the comparison with the external control group. Both the intervention group and the internal control group have low average cholesterol levels in comparison

Figure 76: Fixed effect normal plot of q_i for the external control group comparison of cholesterol levels for men compared with the $N(0,1)$ line



correlation=0.898

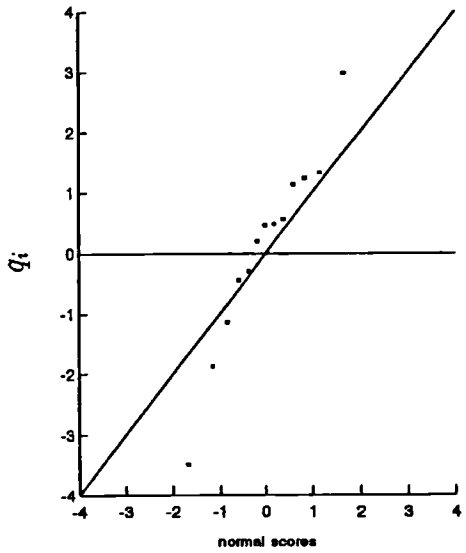
Table 56: Mean levels of blood cholesterol concentration among men for the three study groups in Carlisle

Group	Mean	Standard deviation	Number of men
Intervention	5.438	0.965	180
External control	6.067	1.450	345
Internal control	5.468	1.020	222

to the overall mean level (Table 53), while on the other hand, the external control group has a very high mean cholesterol level (largest individual group mean).

There is heterogeneity present in both comparisons for the cholesterol levels in women but, as would be expected, there is more for the external control group comparison. For the internal control group comparison, there are apparently two outlying points and the rest follow the fixed effect normal model reasonably well (Figure 77). The plot, although not the test (A^2=1.28, p >0.15), suggests that the normally distributed random effects model is not a particularly good fit to the data due to the heterogeneity present.

Figure 77: Fixed effect normal plot of q_i for the internal control group comparison of cholesterol levels for women compared with the $N(0,1)$ line

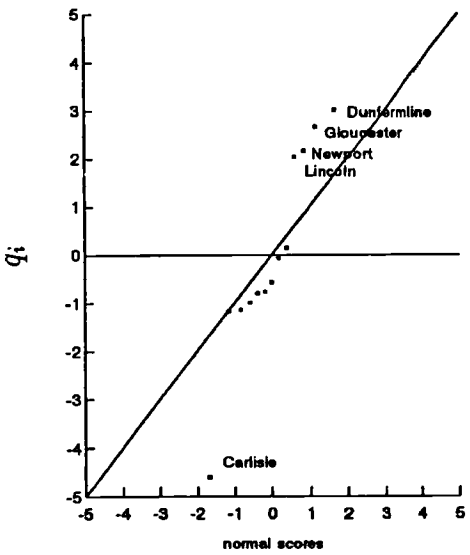


correlation=0.970

For the external control group comparison there is a greater amount of heterogeneity which is spread throughout the 13 towns (Figure 78). There is a group of 4 towns (Gloucester, Lincoln, Dunfermline and Newport) which are similar to each other but noticeably different from the rest of the observed q_i (Figure 78). They all have large positive q , thus indicating that $\hat{\theta}_i > \hat{\theta}$. In fact, in all these towns the

cholesterol level is lower in the external control group than in the intervention group. There is also a clear outlying point with a large negative value of q_i, and, as was the case for the cholesterol level results for men for the external control group comparison, this outlier was Carlisle. Again this is due to a high mean cholesterol level in the control group (largest individual group mean) and a relatively small mean cholesterol level in the intervention group (Table 57). A normally distributed random effects model would also appear to be inappropriate as the groupings observed on the fixed effect plot are still apparent on the random effects plot (Figure 79) and the test for normality is significant (A^2=6.71, p <0.01).

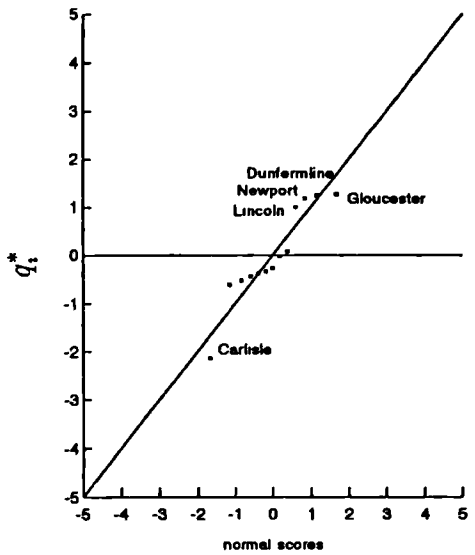
Figure 78: Fixed effect normal plot of q_i for the external control group comparison of cholesterol levels for women compared with the $N(0,1)$ line



correlation=0.946

The heterogeneity in two out of the three examples displaying a significant amount of heterogeneity appears to be due to a few outlying observations rather than being spread throughout the data. The external control group in Carlisle can be singled out as a possibly 'odd' observation contributing to heterogeneity for the results for both men and women. This finding was discussed with the research nurse

Figure 79: Random effects normal plot of q_i^* for the external control group comparison of cholesterol levels for women compared with the $N(0,1)$ line



correlation=0.941

coordinator of the British family heart study, who was familiar with all the participating general practices. It transpired that both practices in Carlisle were situated in the town centre, drawing on similar populations and hence a difference in population characteristics would seem an unlikely explanation for the difference observed. Furthermore, since there was little difference between the mean levels in the internal control group and those in the intervention group, it would not appear that the finding was the result of a particularly effective intervention. A feasible explanation, resulting from the discussion, was that there could be some consistent measurement difference between the intervention and control practice. It is possible that the calibration of the reflotron machines used to measure cholesterol could have been different in each practice, despite a centrally coordinated quality control program [112], so that consistently high readings were obtained in the control practice. It is also possible that the nurse in the external control practice could have made systematically higher readings than the nurse in the intervention group. These theories could be checked

to some degree if the relevant data regarding the calibration of the instrument and measurement of cholesterol were available.

Table 57: Mean levels of blood cholesterol concentration among women for the three study groups in Carlisle

Group	Mean	Standard deviation	Number of women
Intervention	5.249	1.045	109
External control	6.078	1.465	186
Internal control	5.645	1.300	118

Discussion of the large amount of heterogeneity for the results for women in the external control group comparison produced no conclusions. There is no obvious common factor linking the four towns of Gloucester, Lincoln, Newport and Dunfermline, which produce similar residuals in this example. It was only possible to identify two of these towns as possessing unusual characteristics which could explain the results observed, that is the control groups having lower mean cholesterol levels than the intervention groups in these particular town. In one of these towns the two practices were serving populations with difference characteristics, and in the other there were problems with the implementation of the intervention programme. However, these are not convincing as explanations for the heterogeneity since each is related to only one particular town. Furthermore, the two towns have not been identified consistently across different risk factor outcomes (smoking, SBP, DBP and BMI) as having a particularly poor intervention effect and, if there were a true population effect, it would be expected to influence more than one outcome measure.

6.3.3 Analysis of current cigarette smoking

The difference in the prevalence of current cigarette smoking between the intervention group and the control group is now expressed in terms of the log odds ratio, rather than the difference in prevalence rates as in Table 53, and an analysis of this outcome measure is carried out. This is so that the analysis of a binary outcome remains consistent with the measure used throughout the rest of the thesis.

For the internal control group comparisons for both men and women (Tables 58 and 59) all three methods produce the same results since the estimate of the between-town variance for the two random effects methods is zero. For both sexes the results indicate a significantly lower rate of smoking in the intervention group than in the control group, again indicating a benefit from the intervention process.

For both men and women for the external control group comparison, the random effects estimate, using the D&L moment estimator of the between-town variance, of the log odds ratio for the overall intervention effect $\hat{\theta}$ is smaller than the fixed effect estimate, indicating a slightly greater effect due to intervention (Tables 58 and 59). However, the corresponding widening of the confidence interval for θ indicates the reduced certainty in the intervention effect. For men there still remains a significant difference in the smoking rates between the two groups when the heterogeneity is taken into account, with the intervention group having a lower rate than the control group (Table 58). However, for women the confidence interval for the random effects model includes zero whereas for the fixed effect model it does not (Table 59). Hence, the choice of model here affects the conclusions which may be drawn from the analysis with regards to the benefit of intervention on reduction in smoking rates. The random effects likelihood model produces a similar estimate to the standard method of θ for both men and women, but the estimate of the between-town variance is smaller in both cases than the D&L moment estimator. Furthermore, the confidence intervals for θ are, as expected, slightly wider than those derived from the standard random

effects method (Tables 58 and 59), thus further reinforcing the possibility of there being no intervention effect on the prevalence of cigarette smoking among women in the study.

Table 58: Comparison of the results from three different meta-analysis methods for log odds ratios of the prevalence of cigarette smoking comparing intervention and control groups in men in the British family heart study

Comparison group	Method	Estimated between-study variance ($\hat{\sigma}_B^2$)	95% C.I. for σ_B^2	Estimated overall effect ($\hat{\theta}$)	95% C.I. for θ
Internal	Fixed	-	-	-0.242	(-0.399,-0.085)
	Random	0.000	-	-0.242	(-0.399,-0.085)
	Likelihood	0.000	(0.000,0.020)	-0.242	(-0.399,-0.085)
External	Fixed	-	-	-0.197	(-0.342,-0.051)
	Random	0.079	-	-0.228	(-0.442,-0.014)
	Likelihood	0.064	(0.000,0.272)	-0.225	(-0.456,-0.013)

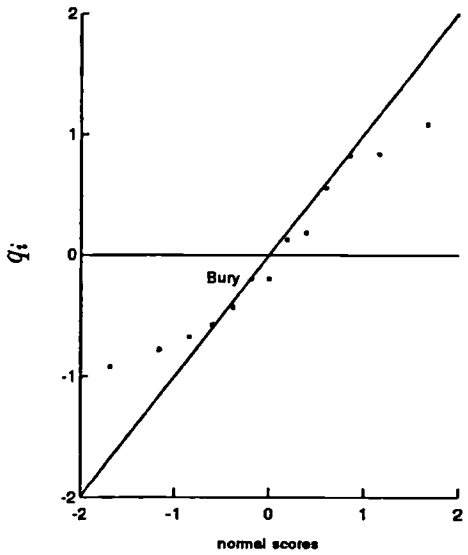
Table 59: Comparison of the results from three different meta-analysis methods for log odds ratios of the prevalence of cigarette smoking comparing intervention and control groups in women in the British family heart study

Comparison group	Method	Estimated between-study variance ($\hat{\sigma}_B^2$)	95% C.I. for σ_B^2	Estimated overall effect ($\hat{\theta}$)	95% C.I. for θ
Internal	Fixed	-	-	-0.227	(-0.427,-0.027)
	Random	0.000	-	-0.227	(-0.427,-0.027)
	Likelihood	0.000	(0.000,0.291)	-0.227	(-0.427,-0.027)
External	Fixed	-	-	-0.198	(-0.382,-0.014)
	Random	0.098	-	-0.213	(-0.469,0.043)
	Likelihood	0.082	(0.000,0.386)	-0.211	(-0.485,0.050)

Fixed effect q-q plots (Section 3.1.1) for the internal control group comparison for both men and women confirm that the results for smoking status are homogeneous across towns as there are no large absolute values of q_i (Figures 80 and 81). For men

the line has a gradient which is less than one, suggesting that there is, in fact, less variation in the estimates than would be expected, while for women the line is not particularly straight. However, a normal fixed effect model would seem satisfactory, although perhaps not ideal, for both these sets of data, as there is certainly no evidence of heterogeneity. Furthermore, the results of the Anderson-Darling test produce test statistics A^2 corresponding to p >0.15 for both men and women.

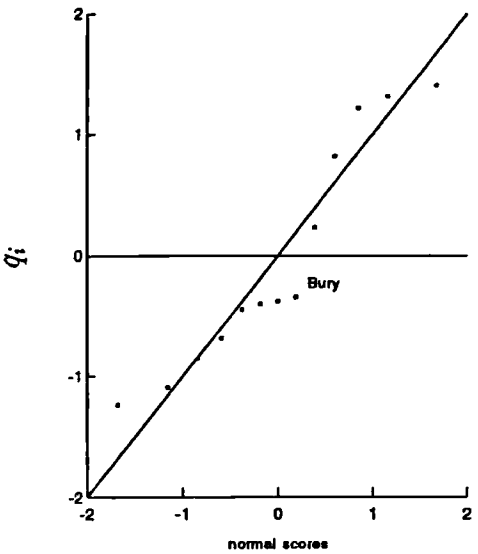
Figure 80: Fixed effect normal plot of q_i for the internal control group comparison of smoking prevalence for men compared with the $N(0,1)$ line



correlation=0.977

The two fixed effect plots for the external control group comparisons (Figures 82 and 83) do indicate the presence of heterogeneity. The plots for men and women are, however, different in that the heterogeneity takes different forms. For men the plot is a straight line through the origin, but with a gradient slightly steeper than unity (Figure 82). Hence, this is consistent with a normally distributed random effects model with reasonably equal variances. Since it is known that the variances of the individual town estimates are fairly similar for this outcome in men (Figure 72), such an interpretation of the plot seems reasonable. The conclusion is further backed up

Figure 81: Fixed effect normal plot of q_i for the internal control group comparison of smoking prevalence for women compared with the $N(0,1)$ line

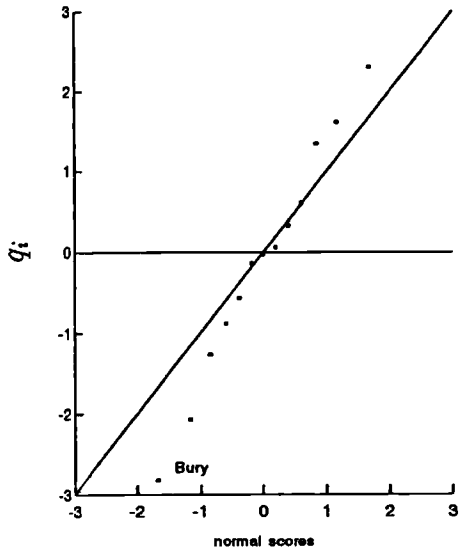


correlation=0.954

by the corresponding random effects plot (Figure 84), which produces a reasonable straight line with a gradient of one. The tests are again of little use here as they appear to lack power in that they find the data consistent with both the fixed effect model and the random effects model. In contrast, for women the fixed effect plot indicates that there is a single very clear outlier which is the sole cause of heterogeneity (Figure 83). All the other points lie along the line of identity indicating that the majority of the town estimates are consistent with each other. On the random effects plot (Figure 85) the outlier has been pulled in, since it has a relatively small within-town variance, but a normally distributed random effects model is less convincing in this situation as backed up by the result of the Anderson-Darling test (A^2=2.29, p <0.1).

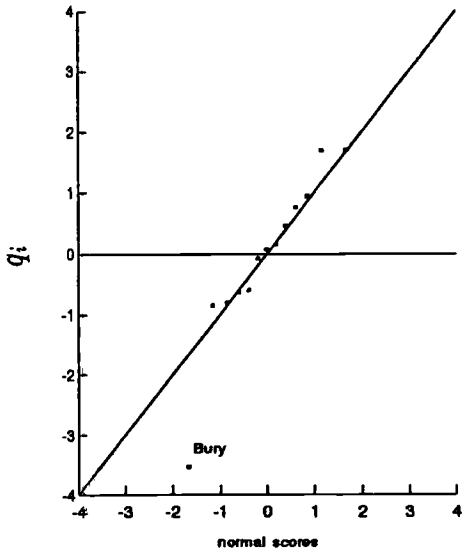
Bury (town number 9) is the town which is identified as a clear outlier in the external control group comparison for women. This town appears to be different with respect to cigarette smoking to all th other towns, especially since it also produces one of the large residuals on the plot for men (Figure 83). As Bury is not noticeably

Figure 82: Fixed effect normal plot of q_i for the external control group comparison of smoking prevalence for men compared with the $N(0,1)$ line



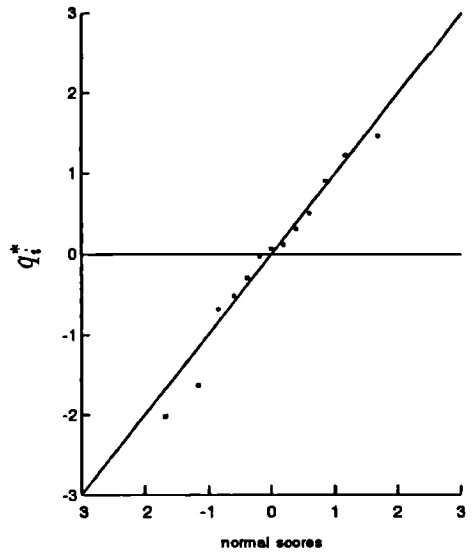
correlation=0.994

Figure 83: Fixed effect normal plot of q_i for the external control group comparison of smoking prevalence for women compared with the $N(0,1)$ line



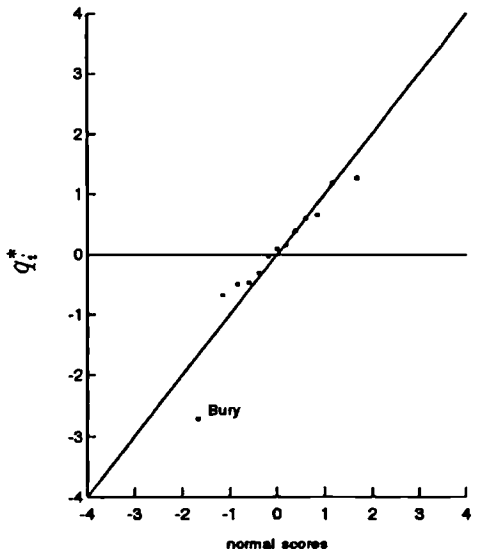
correlation=0.937

Figure 84: Random effects normal plot of q_i^* for the external control group comparison of smoking prevalence for men compared with the $N(0,1)$ line



correlation=0.983

Figure 85: Random effects normal plot of q_i^* for the external control group comparison of smoking prevalence for women compared with the $N(0,1)$ line



correlation=0.930

outlying for the internal control group comparisons (Figure 80 and 81), the suggestion is that the external control group produces the unusual observation. In the external control group in Bury, there were a high proportion of current cigarette smokers, particularly among women (33%), but also among men (30%). Hence, the differences between the smoking rate in the external control group and the intervention group are large, being 17% for men and 23% for women. This is also partly due to the smoking prevalence being rather on the low side in the intervention group, as it is in the internal control group. It should be noted that 30% is actually closer to the national average smoking rate, and hence, it may actually be the other groups where the rates are unusually low and may be as a consequence of a general under reporting of cigarette smoking in the study or of bias caused by the non-randomness of non-respondents.

In relation to the apparently unusually large intervention effect observed in Bury, there are several possible explanations that could be investigated and these were also discussed in the meeting with the research nurse coordinator. Although practices within each town were matched for sociodemographic factors, in some towns the practices were still situated in different areas with differing characteristics and thus drawing on different populations. Bury was, in fact, one such town where there was a substantial geographic difference between the intervention and the control practices, with the intervention practice being in a more advantaged and affluent area than the control practice. Hence, this could explain the difference seen in that a more advantaged population would be expected to have a lower smoking rate. However, greater faith would be placed in this explanation if the same effect were seen in all other towns where there was a population difference between the practices (i.e. that the practice with the more advantaged population always had the lower smoking rate in the control group). This theory could be checked if the relevant data regarding social differences in the practices in each town were available, although such information would be based on rather subjective judgement.

The fact that the cigarette smoking prevalence was low in the internal control group as well as the intervention group in Bury suggests that the large difference in rates is not due to a particularly effective intervention with regards to smoking. However, it is possible that it is due to a difference in reporting rates. Since it is highly unlikely that smoking rates will be overestimated, it may be that in the intervention practice the smoking rates reported were lower than the true rates.

It was also considered possible that the response rate in the control practice could have caused the high prevalence observed. If the response rate was particularly high for the control practice in Bury, then it could be that more smokers responded to the screening than in other centres. However, the response rate was actually about average in this practice and so this would seem an unlikely explanation. Hence, again no convincing explanation for the heterogeneity was found.

6.4 Discussion

Sources of heterogeneity in a meta-analysis require investigation and, therefore, so do the sources of heterogeneity in a paired cluster randomised trial such as the British family heart study. However, because there is more scope for variation between trial protocols in a meta-analysis than in a single multicentre trial where all centres are actually following the same protocol, there may often be more heterogeneity in a meta-analysis than in a multicentre trial. This greater clinical homogeneity may mean that it is more problematic to deduce the causes of heterogeneity in a multicentre trial since there will be fewer and less obvious reasons for the heterogeneity than there would be in a meta-analysis.

In the British family heart study the fact that two control groups were available, in addition to the results for both men and women, meant that there appeared to be a greater chance of sorting out causes of heterogeneity, as all four analyses could

be compared for consistency for each outcome. Hypothetically, if all four comparisons were to show the same unusual effect in any particular town then it would seem likely that this effect would be due to characteristics of that town. If, however, the effect were only seen in one of the control group comparisons, but consistent for both sexes, then the suggestion would be that there is a practice effect present in that town. If, however, the effect were only seen in a single comparison in the town, then the reason would be difficult to deduce, and would probably remain unexplained. Furthermore, if a town were to be identified as unusual for more than one outcome measure then more confidence could be placed in the possibility of a population or town effect.

As has been shown with the analysis of the British family heart study, however, there may not be any obvious or convincing explanations for the heterogeneity observed or for outlying values. For the outcomes considered, that is smoking rates and cholesterol level (Section 6.3) (but also SBP and DBP which are not presented in detail here), no town revealed itself as a consistent outlier across more than one outcome. Hence, this appears to provide evidence against the possibility of differences in patient populations within towns causing different results.

It was only in the case of blood cholesterol levels where the external control group in Carlisle produced outlying results that there was any possibility of a convincing explanation being found, in that it is likely that a consistent calibration or measurement difference was present. In all other cases speculative suggestions are all that appear possible and these are all post-hoc and lacking in any sort of consistency.

Hence, in such situations a random effects analysis may be the best or only solution, if an estimate of an overall intervention effect is required. A fixed effect estimate is inappropriate since the narrow confidence intervals would not reflect the additional uncertainty caused by the between-town variation. So, although investigating sources of heterogeneity rather than resorting to random effects meta-analysis may be commendable in principle, it is not practical in all circumstances, as exempli-

fied here. Furthermore, in general, any such investigation of heterogeneity is post-hoc and so explanations of heterogeneity are based on what is observed in the data. Investigations of heterogeneity may be more problematic in a single multicentre trial than in an actual meta-analysis since there are less clinical differences between centres than there are between separate trials. Separate trials may vary in terms of, for example, patient characteristics, duration of trial and treatment regimen, which may all offer explanations for the variation in individual trial estimates. The problem in a meta-analysis, however, may be that there are too many possible explanations of the variation in the individual trial estimates because the trials vary in many different ways. Hence, there may be a danger that the process of looking too hard for an explanation produces one which is incorrect. Explanations which were put forward as possible causes of heterogeneity before the data were looked at are probably the most reliable, particularly where they are based on sound clinical reasoning. However, a further issue to consider is whether a normally distributed random effects model, such as those used in this analysis, is appropriate. Certainly the normal plots for the examples from the British family heart study do not generally support normality. Hence, there should be some concern over the validity of the results since it is not known how robust the analysis is to deviations from the assumed model.

6.5 Multivariate Models For Meta-Analysis

In a meta-analysis or a paired cluster randomised trial there will often be more than one outcome of interest to be considered, as there is in the British family heart study (Section 6.3). Therefore, as with any single clinical trial with more than one endpoint, the problem of multiple testing and estimation exists and this, therefore, implies that there will be an increase in the overall Type-I error rate. Furthermore, endpoints will usually be correlated with each other and therefore will not be independent. The issue of multiple testing has been considered in the context of meta-analysis by Hedges

and Olkin [39] and Raudenbush et al. [113]. Several possible ways of dealing with multiple endpoints have thus been proposed and discussed. Section 6.5.1 considers simple solutions to the problem by looking at ways of maintaining the correct Type-I error rate. Section 6.5.2 introduces a multivariate approach to meta-analysis and Section 6.5.3 considers the construction of a global test statistic for a multivariate model. The British family heart study is then used as an example in Section 6.5.4 to illustrate the multivariate methods and Section 6.5.5 contains a discussion.

6.5.1 Simple solutions

Hedges and Olkin [39] suggest as one possibility the analysis of only a single endpoint. However, they do acknowledge that this procedure is obviously wasteful of information. This approach is also discussed by Pocock et al. [114] when considering multiple endpoints in a single clinical trial. They suggest the specification of a single primary endpoint in the study protocol which is to be tested formally, with all other endpoints being considered as secondary with the interpretation being exploratory. Such a statement of the primary endpoint of interest could also be made before a meta-analysis is carried out, so that emphasis of the results is focused on the one pre-specified outcome, as opposed to the one which provides the apparently most interesting result. However, there may be situations in which there are several outcomes of equal importance and the discarding or down-weighting of important information is not reasonable.

An alternative suggested by Hedges and Olkin [39] is that each outcome be treated as independent and then the significance levels be adjusted. This is precisely what is sometimes practised in the context of a single clinical trial with multiple endpoints where the individual p-values are modified in order that the overall Type-I error remains at the desired level of α. For example, the Bonferroni inequality can be used for significance tests on p endpoints [115, 116], although this adjustment is

always conservative. The adjustment leads to a nominal significance level for each test being taken as α' where $\alpha' = \alpha/p$. However, when endpoints are correlated, as they tend to be in practice, the Bonferroni correction becomes even more conservative [114]. Furthermore, this solution only applies to hypothesis testing but does not help with regards to estimation.

Raudenbush et al. [113] refer to another strategy for dealing with studies that consider multiple continuous outcome measures in terms of effect sizes. The effect size for study i is given by $\delta_i = (\mu_{i1} - \mu_{i2})/\sigma_{ip}$, where μ_{ij} $(j = 1, 2)$ is the mean in treatment group j and σ_{ip} is the pooled standard deviation. The effect size δ_i is used extensively in psychological research and may be estimated using

$$g_i = (\overline{y}_{i1.} - \overline{y}_{i2.})/\hat{\sigma}_{ip} \tag{128}$$

where $\overline{y}_{ij.}$ is the mean of the individual observations in treatment group j in trial i and

$$\hat{\sigma}_{ip} = \sqrt{\frac{(n_{i1} - 1)s_{i1}^2 + (n_{i2} - 1)s_{i2}^2}{n_{i1} + n_{i2} - 2}} \tag{129}$$

where n_{ij} is the number of observations and s_{ij} is the standard deviation in treatment group j in trial i. However, $g_i = (\overline{y}_{i1.} - \overline{y}_{i2.})/\hat{\sigma}_{ip}$ has been found to be a biased estimate of the true effect size [39] and

$$E(g_i) = \frac{\delta_i}{J(n_{i1} + n_{i2} - 2)} \tag{130}$$

where $J(n_{i1} + n_{i2} - 2)$ may be tabulated [39]. $J(n_{i1} + n_{i2} - 2)$ may be closely approximated by $1 - (3/(4(n_{i1} + n_{i2}) - 9)$ and so

$$E(g_i) \simeq \frac{\delta_i}{1 - \frac{3}{4(n_{i1}+n_{i2})-9}} = \delta_i + \frac{3\delta_i}{4(n_{i1}+n_{i2}) - 9} \tag{131}$$

Hence it can be seen from (131) that, as the sample size becomes large, the bias tends towards zero and so for large trials g_i will be approximately unbiased. The bias may be removed by redefining the estimated effect size so that

$$d_i = J(n_{i1} + n_{i2} - 2)g_i \tag{132}$$

or approximately

$$d_i \simeq \left(\frac{4(n_{i1}+n_{i2})-12}{4(n_{i1}+n_{i2})-9}\right) g_i \tag{133}$$

Therefore, as the sample size increases, the adjustment factor tends to unity.

Hedges and Olkin [39] state that the asymptotic distribution of d_i is normal with mean δ_i and variance

$$var(d_i) = \frac{n_{i1} + n_{i2}}{n_{i1}n_{i2}} + \frac{\delta_i^2}{2(n_{i1} + n_{i2})} \tag{134}$$

The variance may be estimated by replacing δ_i in equation (134) by the estimated effect size d_i.

Effect sizes are therefore standardised measures with no dimensions, which means that endpoints can be combined by taking, for example, a mean or median [117, 118]. This approach is not always possible in medical situations if, for example, the odds ratio is the measurement of treatment effect used, or where an effect size would be difficult to interpret. In many medical trials some sort of combined treatment effect measurement would be meaningless, although it could be used for some psychological or quality of life outcomes.

6.5.2 Multivariate meta-analysis using generalised least squares

An obvious solution to the problem of multiple endpoints is to use multivariate methods in which all outcomes are analysed simultaneously, taking into account the correlations between each pair of outcome measures. Raudenbush et al. [113] propose the use of a generalised least squares (GLS) approach which builds on earlier work by Hedges and Olkin [39] and Rosenthal and Rubin [119]. The model is specified in terms of effect sizes.

Assuming that each trial i, $i = 1, ..., k$ included in the meta-analysis produces results for $r = 1, ..., r_i$ of the total number of p endpoints being considered, the model is of the form

$$\delta = X\beta \tag{135}$$

where δ is the vector of effect sizes $\delta' = (\delta_{11}, \delta_{12}, ..., \delta_{1r_1}, ..., \delta_{k1}, \delta_{k2}, ..., \delta_{kr_k})$ so δ_{ij} is the effect size for the endpoint r for trial i, X is the required design matrix and β is the vector of parameters which are to be estimated. The matrix X has $R = \sum_{i=1}^{k} r_i$ rows, one corresponding to each outcome in each study, with the number of columns being equal to the number of parameters fitted in any particular model. If a single effect is being modelled for all the endpoints then X is a vector containing R 1's. If a different estimate is being obtained for each endpoint, then X has p columns of indicator variables, one for each outcome.

The parameter estimates, together with their variances, may then be obtained using standard GLS techniques. Writing equation (135) in terms of unbiased estimated effect sizes given in (133), the model becomes

$$\mathbf{d} = X\beta + \mathbf{e} \tag{136}$$

where **e** is assumed to be approximately normal with an RxR estimated variance-covariance matrix S. The structure of S is such that it contains the individual trial variance-covariance matrices S_i, $i = 1, .., k$, stacked along the diagonal with all other elements being zero, that is

$$S = \begin{bmatrix} S_1 & 0 & ... & 0 \\ 0 & S_2 & ... & 0 \\ . & . & & . \\ . & . & & . \\ 0 & 0 & ... & S_k \end{bmatrix} \tag{137}$$

Each S_i contains the variance of each effect size in trial i on the diagonal and the covariances of each pair of outcomes elsewhere. The effect sizes for outcomes p and p' have the same asymptotic correlation as the original observations y_{ip} and $y_{ip'}$ [39]. Hence, the covariance between the effect sizes relating to the two correlated outcome measures y_{ip} and $y_{ip'}$ is

$$cov(d_{ip}, d_{ip'}) = \rho_{ipp'}\sqrt{var(d_{ip})var(d_{ip'})} \tag{138}$$

where $\rho_{ipp'}$ is the population correlation between y_{ip} and $y_{ip'}$. In large samples $\rho_{ipp'}$ may be estimated from the sample data within each study.

If **e** is assumed to have a zero mean vector and a known variance-covariance matrix Σ, then the best linear unbiased estimator of β [113] is

$$\hat{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\mathbf{d} \tag{139}$$

and the variance of $\hat{\beta}$ is

$$var(\hat{\beta}) = (X'\Sigma^{-1}X)^{-1} \tag{140}$$

However, since Σ is not known in practice, it must be replaced in equations (139) and (140) by S.

The fit of the model may then be tested by considering the null hypothesis that the predictors included in the model completely explain the variability in effect sizes. The test statistic is given by

$$H_E = (\mathbf{d} - X\hat{\beta})'S^{-1}(\mathbf{d} - X\hat{\beta}) \tag{141}$$

assuming normality for $\mathbf{d}$, and H_E has a χ^2 distribution with $R-q$ degrees of freedom, where q is the number of parameters estimated in the model.

As is usual with regression analysis, the significance of each individual effect $\hat{\beta}_i$ can be tested by considering the null hypothesis $H_0 : \beta_i = 0$ and the familiar z-statistic

$$Z = \frac{\hat{\beta}_i}{\sqrt{var(\hat{\beta}_i)}} \tag{142}$$

is used. Raudenbush et al. [113] then suggest that since each parameter is being considered separately, the p-values associated with each test should be adjusted to avoid an inflated Type-I error probability.

In addition to these individual tests, an overall test of the significance of the model may be carried out. This is a test of whether any of the parameters in the model have a non-zero effect on the outcome and thus is a test of the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$. The test statistic is given by [39]

$$H_R = \mathbf{d}' S^{-1} \mathbf{d} - H_E \tag{143}$$

which has a χ^2_q distribution.

This test H_R is similar to the Hotelling's T^2 test for comparing two multivariate samples and whose test statistic takes the form

$$T^2 = \mathbf{z}' \Lambda^{-1} \mathbf{z} \tag{144}$$

where $\mathbf{z}$ is a vector of z statistics for each of the p endpoints and Λ is the covariance matrix which allows for correlations between the standardised normal deviates. The correlations between the z_i are the same as the correlations between the raw observations. Like Hotelling's T^2 (144) statistic H_R (143) will lack power against certain important alternative hypotheses, since it is a general test of significance looking at whether one or more of the treatment effects are different [120]. The alternative of particular interest is that one treatment performs consistently better than the other for all, or nearly all, of the endpoints. Without the power to detect such an alternative, all the evidence must be weighed up subjectively to deduce which treatment is better overall, rather than being able to give a single probability statement on efficacy. This issue is addressed in Section 6.5.3.

The main problem, however, with this multivariate model is that it is of a fixed effect type and, hence, if any extra variation exists, then the model may not be adequate. In certain circumstances, heterogeneity may be explained by the addition into the model of trial-specific covariates, represented by the addition of a further column to the design matrix. This allows the effect size estimates for each outcome to have a slope, which may be common to all endpoints or different for each endpoint. However, it may sometimes be the case that no covariate can be found which offers a plausible explanation for the heterogeneity. Also, since any such investigation of

covariates is generally post-hoc, caution needs to be expressed. Hence, in some situations the model will be a poor representation of the data and so this approach is limited in its usefulness.

6.5.3 Global test statistic

Addressing the question of a global test statistic which is powerful against specific alternatives, in the context of a single clinical trial, Pocock et al. [114] state that the prime interest is often in an alterative hypothesis with all (or some) endpoints showing treatment differences in the same direction. An alternative to Hotelling's T^2 statistic, which is more powerful for this alternative of interest, was proposed by O'Brien [120]. Assuming $\mathbf{J}' = (1, ..., 1)$, Λ is again the covariance matrix for the multiple endpoints $r = 1, ..., p$, and $\mathbf{z}$ is the vector of test statistics for each individual outcome measure, then, for any p asymptotic normal statistics with known covariance matrix, $\mathbf{J}'\Lambda^{-1}\mathbf{z}$ is the optimal linear combination for the alternative hypothesis that the p standardised treatment differences are all of equal magnitude and in the same direction [114], that is $H_1: z_1 = ... = z_p \neq 0$. The test statistic is therefore of the form

$$\frac{\mathbf{J}'\Lambda^{-1}\mathbf{z}}{(\mathbf{J}'\Lambda^{-1}\mathbf{J})^{1/2}} \sim N(0,1) \tag{145}$$

The weighting factors $\mathbf{J}'\Lambda^{-1}$ are column sums for each variable, indicating their total correlation with all other endpoints. Hence, because Λ is inverted, the endpoints which are less highly correlated with any of the other variables have greater weight. Further work has been carried out which shows how the global test statistic (145) may be extended to any set of asymptotically normal test statistics whose covariance matrix it is possible to estimate [114].

If the same effect size were being assumed for each endpoint in the multivariate meta-analysis model, then the test statistic H_R is equivalent to Hotelling's T^2 since

X becomes **J**, that is a vector of 1's, and each d_{ir} is a standardised difference. The difference with the meta-analysis case is that the null hypothesis being tested is $H_0 : d_{ir} = 0$ for all trials $i = 1, ..., k$ and all outcomes $r = 1, ..., r_i$, rather than $H_0 : z_r = 0$ for all outcomes $r = 1, ..., p$ in a single trial.

The same null hypothesis as for the T^2 test is being tested using the O'Brien type test. The z test (142) in the case where a common treatment effect is being fitted to all outcomes produces a global statistic equivalent to the O'Brien statistic, using (139) and (140) to define $\hat{\beta}$ and $var(\hat{\beta})$. The statistic is given by

$$\frac{\hat{\beta}}{\sqrt{var(\hat{\beta})}} = \frac{X'\Sigma^{-1}\mathbf{d}}{(X'\Sigma^{-1}X)^{1/2}} \tag{146}$$

where X is equivalent to **J** and Σ is also a covariance matrix like Λ and **d** is a vector of asymptotically normal test statistics. This test is more powerful than H_R against the alternative hypothesis of $H_1 : \delta_{11} = \delta_{12} = ... = \delta_{1r_1} = ... = \delta_{k1} = \delta_{k2} = ... = \delta_{kr_k} \neq 0$ and therefore for alternative hypotheses where the effect sizes tend to go in the same direction.

6.5.4 Example of multivariate meta-analysis

The British family heart study is used as an example to illustrate the multivariate methods described in Section 6.5.2. The outcomes considered are diastolic blood pressure (DBP) and systolic blood pressure (SBP). These outcomes were chosen since the value of SBP and DBP for each person will obviously be correlated, and it may also be of interest to make an efficacy statement about the treatment of blood pressure in general, rather than about the two components separately. The correlations could be estimated from the data within each centre and here the correlations for this analysis were actually calculated using both intervention and control observations combined.

The methods have been carried out for the results for men, using both the internal and external control group comparisons. By looking at both control group comparisons, models representing data with differing amounts of heterogeneity may be considered.

Firstly, for each analysis a common difference was fitted for DBP and SBP. Such a model is only reasonable when dealing with effect sizes since it would be meaningless to try and fit the same effect to the mean differences for DBP and SBP, as the differences in SBP are much larger than those in DBP. Secondly, a separate effect was fitted for each blood pressure outcome separately. The fit of both these models was assessed using H_E as defined in (141). The results obtained from such a model could then be compared with the standard individual fixed effect meta-analysis results.

In the univariate analyses, there is less heterogeneity present in the SBP estimates for the internal control group comparison than for the external. However, perhaps surprisingly, there is no evidence of heterogeneity in the DBP estimates for the external control group comparison, but there is some for the internal control group comparison. However, there is overall far less variation in DBP than in SBP and so the multivariate model, being a fixed effect method, is therefore more reasonable for the internal control group comparison. Results for the multivariate model (Table 60) indicate that both blood pressure measurements cannot be adequately estimated by a common effect size for either the internal control group or the external control group comparison. Fitting a separate effect for each outcome does improve the fit of the model significantly in both cases, although neither explains an adequate amount of the variation.

The effect size estimates for both outcomes in both comparisons are very comparable to those obtained from separate univariate fixed effect meta-analyses using effect sizes (Table 61). The standard errors of the two sets of estimates are almost exactly the same in the two cases for both outcomes as well. Hence, nothing

has apparently been gained in this example by carrying out a multivariate meta-analysis

Since in the external control group comparison there is unexplained variation present, because the model is still not a good fit to the data as indicated by the large deviance H_E, one option is to add covariates to the model to try and improve the fit. As an example, purely to illustrate the methodology, a covariate indicating the location of each study centre in Britain was constructed. Since there was one intervention and one control practice situated in 13 towns distributed throughout Britain, a covariate was constructed having two categories indicating a location in either the north (Scotland, north of England and Midlands) or south (south of England, Wales and East Anglia) of the country. A model was then fitted with a common slope for both outcomes and then extended to allow the slope to differ for each of SBP and DBP. In neither of the examples did geographical location explain a significant amount of the variation (Table 60). Hence, there was no geographical variation in the difference in blood pressure between intervention and control groups. It may be the case that any geographical variations have been cancelled by considering differences within towns.

6.5.5 Discussion

In terms of estimation there is little to be gained from the multivariate analysis of the blood pressure outcomes in the British family heart study. However, in other situations where the multiple outcomes of interest may be of equal magnitude the analysis may be worthwhile, as an estimate of a single overall effect size, representing a general treatment effect, could be obtained. The fact that there is no improvement in the precision of the estimates of treatment effect with the multivariate model in the example considered, as opposed to the individual univariate analyses, is because the data is complete, that is there is a measurement recorded for each of DBP and

Table 60: Multivariate generalised least squares models for effect sizes for the difference in blood pressure (both DBP and SBP) between the intervention and control groups in men in the British family heart study

Model	Parameter	Internal control			External control		
		Estimate	S.E.	H_E(df)	Estimate	S.E.	H_E(df)
Common effect size	β	-0.369	0.0297	42.66(25)	-0.292	0.0274	189.10(25)
Separate effect sizes	β_1	-0.336	0.0322	36.05(24)	-0.205	0.0299	135.95(24)
	β_2	-0.403	0.0326		-0.380	0.0298	
Common slope	β_1	-0.369	0.0435	34.82(23)	-0.195	0.0398	135.81(23)
	β_2	-0.435	0.0438		-0.369	0.0397	
	γ	0.066	0.0594		-0.369	0.0548	
Different slopes	β_1	-0.355	0.0451	33.49(22)	-0.249	0.0415	125.46(22)
	β_2	-0.450	0.0456		-0.350	0.0410	
	γ_1	0.037	0.0646		0.091	0.0598	
	γ_2	0.096	0.0650		-0.063	0.0597	

β=common effect size for both DBP and SBP

β_1=effect size for DBP

β_2=effect size for SBP

γ=common slope for both DBP and SBP

γ_1=slope for DBP

γ_2=slope for SBP

H_E=Test of fit of the model given in (141) compared with χ^2_{R-q} distribution

Table 61: Fixed effect meta-analysis results for the difference in blood pressure (both DBP and SBP) between the intervention and control groups in men in the family heart study

Measurement	Control group	DBP		SBP	
		Estimate	S.E.	Estimate	S.E.
Effect size	Internal	-0.336	0.0323	-0.404	0.0324
	External	-0.386	0.0297	-0.207	0.0296
Difference in means	Internal	-3.501	0.514	-7.255	0.568
	External	-2.477	0.568	-7.285	0.514

SBP in each individual and in each town. The multivariate model can cope with different numbers of outcomes being measured in different trials in a meta-analysis. Hence, multivariate methods would provide improvements in precision when different trials measure different outcomes, since such methods make up for missing data in one variable by using the information regarding the others and the correlations between them. This may be useful in certain meta-analyses, where different trials on the same treatment may consider different measures of outcome. It is likely to be less useful for the analysis of multicentre trials because all centres should be measuring the same outcomes as they are following the same protocol. In a meta-analysis these methods would allow the results of more trials to be included, since in order to perform a univariate meta-analysis, all outcome measures must be the same.

Furthermore, the model does, however, provide a global test of treatment effectiveness on all outcomes of interest, which is of some use in that it provides a test of the general impact of the treatment. The test given in (146) may be particularly useful in testing for a consistent treatment effect across all outcomes. Hence, a significant result from such a test provides strong evidence of an overall benefit of treatment.

The fact that the method pursued here is restricted to the use with effect sizes means that interpretation in practice is difficult. With further work it may be possible to adapt the model to cope with alternative measures of treatment effect. Furthermore, although other covariates may explain the heterogeneity present in the blood pressure data, a random effects model may be required to model the additional variation which was evident after the fixed effect model was fitted. Further research is again needed into the issue of fitting a multivariate model which allows for random effects, with a possible way of proceeding being to follow a multilevel modelling approach [121]. This sort of approach is an improvement over the model used here as it allows the covariance structure of the data to be estimated simultaneously with the treatment effect. In the method applied to the blood pressure outcome, the covariance matrix is obtained by estimating the relevant variances and correlations from the data set and then substituting them in to the model as if they were known values. In general, the work illustrated here provides an explanation of the problem of multiple testing in meta-analyses and an introduction to methods which could be pursued and the problems which need to be solved, but do not provide satisfactory solutions for the analysis of the British family heart study.

7 A Comparison of Meta-Analysis and Paired Cluster Randomised Methods

The comparison of the intervention group with the external control group in the British family heart study involves the analysis of a paired cluster randomised design and Section 6.2 illustrated how these data could be analysed using meta-analysis techniques. However, there is also some existing literature which proposes methods specifically for the analysis of paired cluster randomised trials. Hence, in this chapter a comparison of meta-analysis methods and these other paired cluster randomised methods will be made. Both testing and estimation are considered.

Section 7.1 introduces the concept of the intracluster correlation, which is central to paired cluster randomised trial methods. The issue of testing for an overall treatment effect is then considered and existing methods for dichotomous outcomes are described in Section 7.2 and for continuous outcomes in Section 7.3. A discussion of findings in published papers in Section 7.4 is then followed by a comparison of the tests using data from the British family heart study in Section 7.5. Estimation of an overall treatment effect is the focus for the remainder of the chapter, with published methods being described for dichotomous outcomes in Section 7.6 and continuous outcomes in Section 7.7. Again the British family heart study provides an example for the comparison of the different methods and Section 7.8 also compares these methods with the standard meta-analysis methods of Chapter 1. The chapter is rounded off with a discussion and comparison of the different methods (Section 7.9) and a conclusion (Section 7.10).

7.1 Intracluster Correlation

Paired cluster randomised methods use the concept of intracluster correlation. If clusters, rather than individuals, have been randomised, then individuals within a cluster will tend to be more like each other than like individuals from other clusters. This means that the measurements within a cluster are likely to be dependent. The intracluster correlation is a measure of the association between the observations within a cluster compared to between clusters. It may be expressed in terms of the proportion of the total variance due to the between-cluster differences,

$$\rho = \frac{\hat{\sigma}^2_{BC}}{\sigma^2_w + \hat{\sigma}^2_{BC}} \tag{147}$$

where $\hat{\sigma}^2_{BC}$ is the between-cluster variance and σ^2_w is some average within-cluster variance (see Section 7.2). A suitable estimate of this intracluster correlation must be obtained in order to proceed with the methods. The expression for the intracluster correlation is such that if ρ=1, then all the observations in each cluster are exactly the same, that is there is no within-cluster variation. On the other hand if ρ=0, there is no clustering, that is there is no between-cluster variation, and so the observations within a cluster are no more like each other than observations from different clusters. The latter situation is simply a case of straightforward random sampling.

The intracluster correlation links up with the meta-analysis concept as both are concerned with the appropriate use of the between-cluster and within-cluster components of variance. However, the two methods actually estimate the variation differently. In the meta-analysis case, σ^2_B is estimated after taking out a common treatment effect, whereas in the cluster randomised method the estimate of between-cluster variation $\hat{\sigma}^2_{BC}$ is confounded with the treatment effect. This will be explained further in following sections.

7.2 Published Methods For Testing in Paired Cluster Randomised Trials When the Outcome is Dichotomous

In the case of testing for a treatment effect from a meta-analysis point of view, there is no distinction made between the test used for both homogeneous and heterogeneous data. The Mantel-Haenszel test, and its equivalent tests, test the null hypothesis that each individual treatment effect θ_i is zero against the alternative that at least one such effect is non-zero (Section 1.4). They do not test the hypothesis that the overall treatment effect θ is zero, unless homogeneity can be assumed. If homogeneity can be assumed then the Mantel-Haenszel test is not only valid but is also optimal for such a null hypothesis [38].

From the perspective of a paired cluster randomised design, however, the desire to test the null hypothesis that the overall treatment effect is zero when heterogeneity is present, that is $H_0 : \theta = 0$ against the alternative $H_1 : \theta \neq 0$, has led to the development of a range of alternative tests. When positive intracluster correlation is observed in a set of data, individuals within each cluster cannot be regarded as independent and the statistical importance of any one response is decreased. This means that the effective sample size is less than the total number of individuals in the trial, but greater than the total number of clusters (unless there is total dependence within clusters). Hence, due to this effective reduction in sample size, the variance of the treatment effect in each pair of clusters is too small [122]. Consequently, the Mantel-Haenszel test, even when there is no treatment difference, gives significance levels much more extreme than 0.05 when there is a large amount of correlation between the observations in the same cluster [123]. Thus the use of the Mantel-Haenszel test can result in obtaining spurious statistical significance.

The design of a paired cluster randomised trial means that each stratum i ($i =$

$1, ..., k$) includes a pair of matched clusters, one randomly allocated to treatment and the other to control or placebo. This section considers dichotomous outcome measures, and the proportions of patients exhibiting a positive outcome are denoted by $\hat{P}_{i1} = a_i/n_{i1}$ and $\hat{P}_{i2} = c_i/n_{i2}$ where $i = 1, ..., k$ (1=intervention, 2=control). Five different tests of the hypothesis $H_0 : \theta = 0$ for dichotomous outcome measures, which have been published in the literature, will now be presented in Sections 7.2.1 to 7.2.5. An example illustrating the use of and a comparison of the results of these tests is presented in Section 7.5.

7.2.1 Unweighted t-test

The simple unweighted paired t-test, applied to dichotomous outcome measures, is given by,

$$t_u = \frac{\hat{\theta}_u \sqrt{k}}{\sqrt{\sum_{i=1}^{k} (\hat{\theta}_i - \hat{\theta}_u)^2/(k-1)}} \tag{148}$$

where $\hat{\theta}_u = \sum_{i=1}^{k} \hat{\theta}_i/k$ and $\hat{\theta}_i = \hat{P}_{i1} - \hat{P}_{i2}$ is the estimate of the treatment effect in strata i. Here each cluster mean is treated as if it were a single observation. Hence, this test assumes that all the variation is between-clusters and no account is taken of the variation within each cluster or of the different cluster sizes. Under the null hypothesis and assuming normality of $\hat{\theta}_i$, this statistic has a t distribution with $(k-1)$ degrees of freedom.

7.2.2 Weighted t-test for proportions

Donner and Donald [123] present two alternative approaches for dichotomous data based on a t-statistic which allow for both within-cluster and between-cluster variation. The first method is based on a weighting of the differences $\hat{\theta}_i = \hat{P}_{i1} - \hat{P}_{i2}$. The

weighted average is $\hat{\theta}_w = \sum_{i=1}^{k} w_i\hat{\theta}_i / \sum_{i=1}^{k} w_i$, where an appropriate choice of weights is found by considering the variance of $\hat{\theta}_i$ under the null hypothesis $H_0 : \theta = 0$. If n_{i1} and n_{i2} are reasonably large in all strata, then the variance is approximated by

$$var(\hat{\theta}_i) = \overline{P}_i(1 - \overline{P}_i)\left[\frac{1+(n_{i1}-1)\hat{\rho}}{n_{i1}} + \frac{1+(n_{i2}-1)\hat{\rho}}{n_{i2}}\right] \tag{149}$$

where $\overline{P}_i = (a_i + c_i)/(n_{i1} + n_{i2})$ is the estimated event rate in stratum i under the null hypothesis and $\hat{\rho}$ is an estimate of the intracluster correlation (Section 7.1). As each n_i tends to infinity, then $var(\hat{\theta})$ reduces to $2\hat{\rho}\overline{P}_i(1 - \overline{P}_i)$ approximately. Hence, an estimate of ρ is required to proceed and the derivation of such an estimate is now provided.

Analysis of variance methods are used to obtain $\hat{\rho}$. Unbiased estimates of the average within-cluster correlation σ_w^2 and the between-cluster variation σ_B^2 are given by [124],

$$\hat{\sigma}_w^2 = MSE \tag{150}$$

$$\hat{\sigma}_{BC}^2 = (MSC - MSE)/n_A \tag{151}$$

where MSE is the error mean square, MSC is the cluster mean square and n_A is an adjusted average sample size and is given by $[N - \sum_{i=1}^{k}(\sum_{j=1}^{2} n_{ij}^2/n_i)]/k$. The derivation of the required analysis of variance table for the example of particular interest is provided later. Hence, for now, it is simply stated that the estimate of ρ is given by

$$\hat{\rho} = \frac{\hat{\sigma}_B^2}{\hat{\sigma}_w^2 + \hat{\sigma}_B^2} = \frac{(MSC - MSE)/n_A}{MSE + (MSC - MSE)/n_A} = \frac{MSC - MSE}{MSC + (n_A - 1)MSE} \tag{152}$$

If $MSC < MSE$ then $\hat{\rho}$ must be set to zero as the intracluster correlation cannot sensibly take a negative value. Once an estimate of the intracluster correlation has been obtained, the weights w_{ip}, which are the reciprocal of the variances (equation (149)) which take account of the clustering, can be calculated. The estimate of the overall treatment effect is therefore given by

$$\hat{\theta}_{wp} = \sum_{i=1}^{k} w_{ip}\hat{\theta}_i / \sum_{i=1}^{k} w_{ip} \tag{153}$$

The test statistic for looking at the treatment effect over all strata, allowing for clustering given by Donner and Donald [123] takes the form

$$t_{wp} = \frac{\hat{\theta}_{wp} \sum_{i=1}^{k} w_{ip}}{s_d \sqrt{\sum_{i=1}^{k} {w_{ip}}^2}} \tag{154}$$

where $s_d^2 = \sum_{i=1}^{k} w_{ip}(\hat{\theta}_i - \hat{\theta}_{wp})^2 / \sum_{i=1}^{k} w_{ip}$. Under H_0, t_{wp} has an approximate t-distribution with $(k-1)$ degrees of freedom. If the number of strata is small, Donner and Donald [123] indicate that it is desirable to introduce a continuity correction. This involves replacing $\hat{\theta}_{wp}$ by $\hat{\theta}_{wp} - 0.5/\sum_{i=1}^{k} n_{i1} - 0.5/\sum_{i=1}^{k} n_{i2}$.

No details regarding the derivation of the statistic are, however, provided by Donner and Donald [123]. Hence, to understand the reasoning behind the method a derivation is now provided which has been deduced through reference to other papers.

The estimate is derived from the analysis of variance table used in the analysis of intraclass correlation in multiple samples for survey data [124]. In this multiple sample situation there are c_i clusters within each stratum i, but there are no different 'treatments' within each stratum. The model can therefore be written as

$$y_{ijl} = \mu + \alpha_i + \beta_{ij} + e_{ijl} \tag{155}$$

where $\beta_{ij} \sim N(0, \sigma^2_{Bc})$ and $e_{ijl} \sim N(0, \sigma^2_w)$ with $i = 1, ..., k, j = 1, ..., c_i$ and $l = 1, ..., n_{ij}$. The stratum effects α_i are considered to be fixed, while the cluster effect β_{ij} and the individual effect e_{ijl} are both random. The intracluster correlation is assumed to be constant across all strata, which implies that the variation in the response y from cluster to cluster is the same in each stratum [124].

In adapting this approach to a paired cluster randomised design, a treatment effect is introduced within each stratum, which, however, the model does not account for. When constructing the analysis of variance table required for the model given in equation (155) a stratum mean square and a cluster (within stratum) mean square are obtained, together with the error mean square. Hence, the variation between clusters within each stratum will be due in part to a treatment effect, if one exists. Hence, due to the fact that $\hat{\rho}$ is based on the variation between clusters within a stratum [125], an unbiased estimate of ρ may only be obtained under the null hypothesis of no treatment effect.

To derive the relevant analysis of variance table for such binary outcomes it must initially be assumed that each observation is on a continuous scale but can take one of only two values, either 0 or 1. The value 1 is recorded if an individual exhibits the outcome of interest and otherwise 0 is recorded. Then letting y_{ijl} be the observed value (0 or 1) for individual l (l=1,...,n_{ij}) in cluster j (j=1,2) of strata i (i=1,...k), the analysis of variance table (Table 62) may be constructed [124].

Then since y_{ijl} can only take the values 0 and 1, then $\sum_{l=1}^{n_{ij}} y_{ijl}$ is equal to the number of positive responses, a_{ij} say, in cluster j of stratum i ($a_{i1} = a_i$ and $a_{i2} = c_i$). Hence, the sums of squares in Table 62 may be simplified because,

$$\sum_{l=1}^{n_{ij}} y_{ijl} = a_{ij} \tag{156}$$

Table 62: Analysis of variance table for a paired cluster randomised design

Source of variation	Degrees of freedom	Sums of squares	Mean square
Strata	$k-1$	$SSS = \sum_{i=1}^{k} n_i(\overline{y}_{i..} - \overline{y}_{...})^2$	$MSS = SSS/k-1$
Cluster	k	$SSC = \sum_{i=1}^{k}\sum_{j=1}^{2} n_{ij}(\overline{y}_{ij.} - \overline{y}_{i..})^2$	$MSC = SSC/k$
Error	$N-2k$	$SSE = \sum_{i=1}^{k}\sum_{j=1}^{2}\sum_{l=1}^{n_{ij}}(y_{ijl} - \overline{y}_{ij.})^2$	$MSE = SSE/N-2k$
Total	$N-1$	$SST = \sum_{i=1}^{k}\sum_{j=1}^{2}\sum_{l=1}^{n_{ij}}(y_{ijl} - \overline{y}_{...})^2$	

$\overline{y}_{ij.} = \sum_{l=1}^{n_{ij}} y_{ijl}/n_{ij}$=mean of all individuals belonging to the j^{th} cluster in strata i.

$\overline{y}_{i..} = \sum_{j=1}^{2}\sum_{l=1}^{n_{ij}} y_{ijl}/n_i$=mean of all individuals in strata i.

$\overline{y}_{...} = \sum_{i=1}^{k}\sum_{j=1}^{2}\sum_{l=1}^{n_{ij}} y_{ijl}/N$=mean of all individuals in the study.

---

and

$$\sum_{l=1}^{n_{ij}} y_{ijl}^2 = a_{ij} \tag{157}$$

$$\overline{y}_{i..} = \frac{1}{n_{i1}+n_{i2}}\sum_{j=1}^{2} a_{ij} = \overline{P}_i \tag{158}$$

$$\overline{y}_{ij.} = \frac{a_{ij}}{n_{ij}} = \hat{P}_{ij} \tag{159}$$

Hence,

$$SSC = \sum_{i=1}^{k}\sum_{j=1}^{2} n_{ij}(\hat{P}_{ij} - \overline{P}_i)^2 \tag{160}$$

and

$$
\begin{aligned}
SSE &= \sum_{i=1}^{k}\sum_{j=1}^{2}[\sum_{l=1}^{n_{ij}}(y_{ijl}-\tfrac{a_{ij}}{n_{ij}})^2] \\
&= \sum_{i=1}^{k}\sum_{j=1}^{2}[\sum_{l=1}^{n_{ij}}(y_{ijl}^2-2y_{ijl}\tfrac{a_{ij}}{n_{ij}}+\tfrac{a_{ij}^2}{n_{ij}^2})] \\
&= \sum_{i=1}^{k}\sum_{j=1}^{2}[a_{ij}-2\tfrac{a_{ij}^2}{n_{ij}}+\tfrac{a_{ij}^2}{n_{ij}}] \\
&= \sum_{i=1}^{k}\sum_{j=1}^{2}a_{ij}(1-\tfrac{a_{ij}}{n_{ij}}) \\
&= \sum_{i=1}^{k}\sum_{j=1}^{2}a_{ij}(1-\hat{P}_{ij})
\end{aligned}
\tag{161}
$$

Hence, the values for the relevant mean squares calculated using the expressions in (160) and (161) can be substituted into the expression for the estimate of the intracluster correlation (152). The variance of each $\hat{\theta}_i$ (149) may then be estimated in order to obtain the weights w_{ip}. These weights are then used to obtain the estimate of the overall treatment effect $\hat{\theta}_{wp}$ (153), although they are assumed known rather than estimated.

The test statistic is then given by this overall estimate divided by its standard error, that is $t_{wp} = \hat{\theta}_{wp}/\sqrt{var(\hat{\theta}_{wp})}$. Now equating this expression with that in (154) implies that

$$
\begin{aligned}
var(\hat{\theta}_{wp}) &= \sum_{i=1}^{k} {w_{ip}}^2 s_d^2/(\sum_{i=1}^{k} w_{ip})^2 \\
&= \sum_{i=1}^{k} {w_{ip}}^2 var(\hat{\theta}_i)/(\sum_{i=1}^{k} w_{ip})^2
\end{aligned}
\tag{162}
$$

Hence, Donner and Donald appear to be estimating $var(\hat{\theta}_i)$ by $s_d^2 = \sum_{i=1}^{k} w_{ip}(\hat{\theta}_i - \hat{\theta}_{wp})^2/\sum_{i=1}^{k} w_{ip}$ for all i, rather than estimating each variance separately using (149). Hence, a pooled between-stratum estimate of the variance is used, which would appear incorrect since it is being assumed that all the individual stratum estimates are equal when they are not under clustering.

Donner and Donald [123] do acknowledge a drawback with this method in that the way the model is defined leads to the variation between clusters being combined with the treatment effect. Hence, "as the treatment effect increases, $\hat{\rho}$ will also

increase and will provide an increasingly biased estimate of the true ρ" [126]. Thus the intracluster correlation attributable purely to the design of the study cannot be obtained. This issue is discussed and investigated further in following sections of this chapter.

7.2.3 Empirical logistic weighted t-test

An analogous procedure to the test described in Section 7.2.2 was also presented by Donner and Donald [123] based on the empirical logistic transform,

$$\hat{\theta}_{il} = log\left(\frac{(a_i+0.5)(d_i+0.5)}{(c_i+0.5)(b_i+0.5)}\right).$$

The variance of $\hat{\theta}_{il}$ may be approximated by [123]

$$var(\hat{\theta}_{il}) = \frac{(n_{i1}+1)(n_{i1}+2)[1+(n_{i1}-1)\hat{\rho}]}{n_{i1}(a_i+1)(b_i+1)} + \frac{(n_{i2}+1)(n_{i2}+2)[1+(n_{i2}-1)\hat{\rho}]}{n_{i2}(c_i+1)(d_i+1)} \tag{163}$$

Again, taking a weighted average with the weights w_{il} being equal to the reciprocal of the variance of each estimate (163), a test statistic is obtained in the same way as for the weighted t-test.

The weighted average is therefore $\hat{\theta}_{wl} = \sum_{i=1}^{k} w_{i_l}\hat{\theta}_{il} / \sum_{i=1}^{k} w_{il}$ and in order to test the null hypothesis that there is no overall treatment effect, the test statistic

$$t_{wl} = \frac{\hat{\theta}_{wl}\sum_{i=1}^{k} w_{i_l}}{s_l\sqrt{\sum_{i=1}^{k} w_{il}{}^2}} \tag{164}$$

where $s_l^2 = \sum_{i=1}^{k} w_{i_l}(\hat{\theta}_{il} - \hat{\theta}_{wl})^2 / \sum_{i=1}^{k} w_{il}$, can be compared to the t-distribution with $(k-1)$ degrees of freedom. Hence, similar to the weighted t-test of Section 7.2.2, this leads to the conclusion that $var(\hat{\theta}_{wl}) = \sum_{i=1}^{k} w_{il}s_l^2/(\sum_{i=1}^{k} w_{il})^2$ which means that

$var(\hat{\theta}_i) = s_l^2$, that is a pooled estimate of the variance of $\hat{\theta}_i$.

7.2.4 Wilcoxon signed rank test

The straightforward non-parametric Wilcoxon signed-rank test may also be used in the analysis of a paired cluster randomised trial. Such a test considers the null hypothesis that the overall median treatment difference is zero. It uses both the direction of the difference between a pair of clusters as well as the ranks of these differences, but not their magnitude or precisions. In the situation under discussion this test may be applied to the difference in proportions $\hat{\theta}_i = \hat{P}_{i1} - \hat{P}_{i2}$ for $i = 1, ..., k$.

The $\hat{\theta}_i$ are ranked without regard to their signs. Rank 1 is therefore assigned to the smallest absolute difference and rank k to the largest absolute difference. The test is then based on the value T^+, the sum of the ranks of the positive differences. For small k, the exact distribution of T^+, which is symmetrical about $k(k+1)/4$ can be tabulated. However, for large k ($k \geq 15$) it can be assumed that T^+ is approximately normal. It may be shown [127] that

$$E(T^+) = \frac{k(k+1)}{4} \tag{165}$$

$$var(T^+) \simeq \frac{k(k+1)(2k+1)}{24} \tag{166}$$

Hence, the test statistic

$$Z = \frac{T^+ - E(T^+)}{\sqrt{var(T^+)}} \tag{167}$$

can be compared to a standard normal distribution.

7.2.5 Permutation test

An alternative non-parametric approach is to use a permutation test which uses the magnitudes, rather than simply the directions, of the differences $\hat{\theta}_i = \hat{P}_{i1} - \hat{P}_{i2}$ as well as their ranks, but not their precisions. The rationale behind such a test, is that under the null hypothesis $H_0 : \theta = 0$, the event rates would remain the same if the labels 'treatment' and 'control' within a pair of clusters were interchanged. This is equivalent to regarding the assignment of labels within a pair of clusters as random. This means that, for each strata, the observed difference may be regarded as either positive or negative with equal probability. There are thus 2^k equally likely possible combinations of these signs under the null hypothesis conditional on the magnitudes of the differences actually observed. Hence, 2^k separate differences $D = \sum_{i=1}^{k} \hat{\theta}_i$ can be calculated from the observed data. Therefore a one-sided test rejects at the 5% level if the observed total falls among the largest $0.05 \text{x} 2^k$ values. If k is too small, then $0.05 \text{x} 2^k$ could be less than 1, hence this method would not be very sensible. On the other hand, unless k is reasonably small, this process will be time consuming and, since the distribution of D depends on the difference actually observed, it is not practical to tabulate D. Hence, in practice, approximate forms of the test are used. Donner and Donald [123] use an approximation based on the fact that under $H_0 : \theta = 0$, the mean difference $\hat{\theta}_u = \sum_{i=1}^{k} \hat{\theta}_i / k$ has a normal distribution with zero mean and variance $\sum_{i=1}^{k} (\hat{\theta}_i - 0)^2 / k^2 = \sum_{i=1}^{k} \hat{\theta}_i^{\,2} / k^2$. Hence the test is based on the statistic

$$Z = \frac{\sum_{i=1}^{k} \hat{\theta}_i / k - 0}{\sqrt{\sum_{i=1}^{k} \hat{\theta}_i^{\,2} / k^2}} = \frac{\sum_{i=1}^{k} \hat{\theta}_i}{\sqrt{\sum_{i=1}^{k} \hat{\theta}_i^{\,2}}} \tag{168}$$

which has an approximate standard normal distribution under the null hypothesis $H_0 : \theta = 0$.

Gail et al. [128], use an equivalent test based on an approximation of the test statistic to the t-distribution [129]. The test statistic is the same as that above (168) apart from the fact that a different estimate of the variance is used. The variance of $\hat{\theta}_u$ suggested is given by $\sum_{i=1}^{k}(\hat{\theta}_i - \hat{\theta}_u)^2/(k-1)k = s^2/k$ and so the test statistic is

$$T = \frac{\sum_{i=1}^{k} \hat{\theta}_i/k - 0}{\sqrt{s^2/k}} = \frac{\sum_{i=1}^{k} \hat{\theta}_i}{s\sqrt{k}} \tag{169}$$

7.3 Published Methods For Testing in Paired Cluster Randomised Trials When the Outcome Variable is Continuous

In the case of a continuous outcome measure Donner and Klar [130] propose the use of either an unweighted paired t-test (Section 7.3.1) or a weighted paired t-test (Section 7.3.2). Their paper refers to estimation and confidence intervals, but obviously tests may also be derived and it is these that are presented here. In the case of a continuous outcome variable the estimated treatment effect in strata i is given by the difference in means, $\hat{\theta}_i = \overline{y}_{i1} - \overline{y}_{i2}$. A generalisation of the paired t-test, proposed by Rosner [81], which allows for heterogeneity between strata, is then described in Section 7.3.3.

7.3.1 Unweighted paired t-test

The same straightforward unweighted paired t-test, not taking account of the within-cluster variance, that was used for a dichotomous outcome (Section 7.2.1) may also be used for a continuous outcome. Hence, the overall treatment effect may be estimated as

$$\bar{d} = \frac{\sum_{i=1}^{k} \hat{\theta}_i}{k} \tag{170}$$

which has an estimated variance given by

$$s^2 = \frac{\sum_{i=1}^{k} (\hat{\theta}_i - \bar{d})^2}{(k-1)} \tag{171}$$

The test statistic therefore takes the form

$$t = \frac{\bar{d}}{s\sqrt{k}} \tag{172}$$

which has a t-distribution with $(k-1)$ degrees of freedom.

The unweighted paired t-test is only strictly valid if $n_{i1} = n_{i2} = n$ for $i = 1, ..., k$, because variance homogeneity must be assumed [130]. Hence, when cluster sizes are moderately or severely imbalanced Donner and Klar [130] state that a weighted procedure would be preferable.

7.3.2 Weighted t-test

The weighted t-test uses a weighted average of the individual strata estimates of the treatment effect,

$$\bar{d}_w = \frac{\sum_{i=1}^{k} w_{iw} \hat{\theta}_i}{\sum_{i=1}^{k} w_{iw}} \tag{173}$$

The estimated variance given by Donner and Klar [130] is

$$var(\bar{d}_w) = \frac{s_w^2 \sum_{i=1}^{k} w_{iw}^2}{[\sum_{i=1}^{k} w_{iw}]^2} \tag{174}$$

where $s_w^2 = \sum_{i=1}^k w_{iw}(\hat{\theta}_i - \bar{d}_w)^2 / \sum_{i=1}^k w_{iw}$, again implying that the variance of the individual stratum effects have been obtained using a between-strata estimate, rather than within-strata estimates. Hence a test statistic is given by

$$t_w = \frac{\bar{d}_w}{s_w \sqrt{\sum_{i=1}^k w_{iw}^2 / \sum_{i=1}^k w_{iw}}} = \frac{\bar{d}_w \sum_{i=1}^k w_{iw}}{s_w \sqrt{\sum_{i=1}^k w_{iw}^2}} \tag{175}$$

A reasonable, but simple, choice of weights, according to Donner and Klar [130], is given by $w_{iw} = n_{i1}n_{i2}/(n_{i1} + n_{i2})$. These weights are based solely on sample sizes and therefore do not depend upon the individual cluster variances.

7.3.3 Rosner's generalisation of the paired t-test

Rosner [81] proposed a generalised paired t-test for continuous outcome measures. It is an extension of the standard paired t-test to a situation where there are variable numbers of cases and controls per pairing. This situation is therefore directly applicable to meta-analysis and to the case where each pairing is made up of two clusters.

It is assumed that the within-strata differences between the treatment group and the control group means follow a one-way random effects analysis of variance model,

$$\hat{\theta}_i = \bar{y}_{i1.} - \bar{y}_{i2.} = \theta + \alpha_i + e_i, \quad i = 1, ..., k \tag{176}$$

where θ is the overall within-strata difference in means, α_i is the random effect representing a stratum specific change in difference and $\alpha_i \sim N(0, \sigma_B^2)$. Then e_i is the variation within group for strata i and $e_i \sim N(0, v_i)$ where $v_i = \sigma^2(\frac{1}{n_{i1}} + \frac{1}{n_{i2}})$. This is exactly the model for the normally distributed random effects meta-analysis

(Section 1.7.1).

Once again the aim is to test the null hypothesis $H_0 : \theta = 0$ against the alternative $H_1 : \theta \neq 0$. As in Section 2.2.1, the marginal distribution of each stratum estimate $\hat{\theta}_i$ has a normal distribution with mean θ and variance $(\sigma_B^2 + v_i)$ and so the full likelihood for all the strata may be obtained (Section 2.2.1)

$$L(\theta, \sigma_B^2) = \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi(\sigma_B^2 + v_i)}} exp\left\{ \frac{-(\hat{\theta}_i - \theta)^2}{2(\sigma_B^2 + v_i)} \right\} \tag{177}$$

The variances of the individual estimates are then estimated using the following unbiased estimator for σ^2,

$$\hat{\sigma}^2 = \sum_{i=1}^{k} \sum_{j=1}^{2} \sum_{l=1}^{n_{ij}} (y_{ijl} - \overline{y}_{ij.})^2 / (N - k) \tag{178}$$

Then conditional on σ^2, the maximum likelihood estimates of θ and σ_B^2 are calculated. Two equations are obtained which must be solved iteratively to produce the estimates (Section 2.2.1).

The variance-covariance matrix for the vector $\Phi = (\hat{\theta}_r, \hat{\sigma}_B^2)^T$, given by $I^{*-1}(\hat{\Phi})$, can be found where $I^*(\hat{\Phi})$ is the observed information matrix for $\hat{\theta}$ and $\hat{\sigma}_B^2$ as defined in Section 2.3.4 and w_i=$1/var(\hat{\theta}_i)$ where $1/var(\hat{\theta}_i)$ is estimated individually for each i,

$$I^*(\hat{\Phi}) = \begin{bmatrix} \sum_{i=1}^{k} w_i & \sum_{i=1}^{k} w_i^2(\hat{\theta}_i - \hat{\theta}_r) \\ \sum_{i=1}^{k} w_i^2(\hat{\theta}_i - \hat{\theta}_r) & \sum_{i=1}^{k} w_i^3(\hat{\theta}_i - \hat{\theta}_r)^2 - \sum_{i=1}^{k} w_i^2/2 \end{bmatrix}$$

This is exactly the matrix used in Section 2.3.4 to provide approximate confidence intervals for the maximum likelihood estimates of θ and σ_B^2.

Rosner [81] uses the variance-covariance matrix to obtain an asymptotic test procedure for the null hypothesis of no overall treatment effect. The test statistic is

given by

$$\lambda = \frac{\hat{\theta}_r}{\sqrt{I_{11}^{*-1}(\hat{\theta}_r)}} \tag{179}$$

where $I_{11}^{*-1}(\hat{\theta}_r)$ is entry $(1,1)$ in the 2x2 variance-covariance matrix, and the test compares this with the standard normal distribution.

7.4 Discussion of Existing Methods

Donner and Donald [123] carried out a Monte Carlo investigation to look at the powers and significance levels of tests described in Sections 7.2.1 to 7.2.4 and to compare them with the Mantel-Haenszel test. The quantities k, n_{ij} and P_{ij} were varied under a paired cluster randomised design. The number of strata k was taken to be 6 and then 12. The odds ratio $\psi = P_{i1}(1 - P_{i2})/P_{i2}(1 - P_{i1})$ was fixed at 1 for the null procedure used to obtain significance levels and at 1.5 for the procedure used to compare powers. By fixing ψ and varying the value of P_{i2}, the value of P_{i1} is automatically assigned. For k=6, P_{i2} (i=1,...,6) were taken to be 0.3, 0.5, 0.7, 0.3, 0.5, 0.7 respectively, while for k=12, this pattern was repeated twice. The simplification that $n_{i1} = n_{i2} = n_i$ was then made and three levels of imbalance in numbers from stratum to stratum (balanced – n_i=120 (i=1,...,k), mildly imbalanced – n_i=60, 120, 180, 60, 120,..., severely imbalanced – n_i=20, 120, 220, 20, 120,...) were considered. For each combination of k and balance of design, various values of the intracluster correlation, ranging from 0 to 0.15, were investigated. This was done both for significance level (ψ=1) and for power (ψ=1.5) simulations.

The Mantel-Haenszel procedure gave significance levels much more extreme than 0.05 in situations where ρ was greater than zero, thus indicating the inappropriateness of this test for clustered data when testing $H_0 : \theta = 0$. As the intracluster

correlation increased, the significance level became more extreme.

The standard unweighted paired t-test assumes that there is no within-cluster variability as it considers the rate from each cluster as a single observation. Hence, this test is also strictly inappropriate in the situation where clustering is present. However, the Monte Carlo study found that the paired t-test t_u (148) provided satisfactory significance levels in general for all factor combinations. Donner and Donald therefore suggest that this illustrates "the robustness of this procedure to departures from normality and homogeneity of variance". The homogeneity of variance being referred to is that of the variances of $\hat{\theta}_i$.

The paired t-test may also be applied to a continuous outcome measure for a paired cluster randomised design (Section 7.3.1). However, Rosner [81] indicates that the standard paired t-test is not valid unless there is no within-cluster variation (or the numbers in each cluster are the same). He therefore advocates the use of the generalised paired t-test (Section 7.3.3), but notes that, if σ_B^2/σ_w^2 is large, then the result will be almost exactly that of the standard paired t-test. This is because when the variation within each cluster is very small, compared to the differences between clusters, then the within cluster variation can effectively be ignored and each cluster can be treated as if it provides only a single observation.

Donner and Donald, in the 1987 paper [123], suggest that the standard paired t-test is adequate for all situations where continuous outcomes are involved. This assertion stems from investigations carried out by Korn [131] which indicate that the standard paired t-test can be recommended for most practical situations. Korn shows that the asymptotic relative efficiency of the paired t-test is very high (>0.89 for all cases considered). By a simulation study he also shows that the rejection probabilities of the standard test are around 0.05 even when there is clustering. However, this investigation by Korn is limited and only considers the case where there are small sized clusters, as the specific application under discussion is for a paired case-control

design with differing numbers of cases and controls per strata. Hence, a generalisation to all designs regarding the robustness of the paired t-test from this evidence alone may not be valid. Moreover, Donner appears to have revised his view, as a later paper [130] suggests that the use of a weighted paired t-test is preferable in situations where the cluster sizes are moderately or severely imbalanced.

From the Monte Carlo study [123], Donner and Donald found that for cases where the intracluster correlation was greater than 0.05, the weighted paired t-test and the logistic weighted t-test were generally more powerful than the unweighted test. This was particularly noticeable where the numbers of individuals between strata were severely imbalanced. The weighted logistic t-test was also found to be useful if there was considerable variation in the event rates from stratum to stratum. The non-parametric test was found to be less powerful in every case than all the parametric tests, although the difference was less when k=12 than when k=6.

In conclusion, Donner and Donald suggest the use of a weighted procedure when a design involves a few strata each of a fairly large size and where the intracluster correlation is likely to be small but significant. The weighted logistic test is slightly favoured over the weighted t-test as it produced nominal significance levels closer to 0.05 in the simulation study, while both tests are approximately equal with regards to power. For studies involving a large number of small strata, the standard paired t-test or a non-parametric test is recommended. The advantage in power of the weighted tests over the unweighted test becomes minimal for larger values of k, particularly when the numbers per cluster are reasonably balanced. Donner and Donald state that they do not expect weighted tests to perform well for designs with many small strata. However, they give no argument as to why this should be so, and the results from the simulation study do not back up this view. Presumably, problems may be caused by the necessity of estimating many weights imprecisely (as exemplified in Chapter 5). It is clear that the advantage in power of the weighted procedures is diminished in

such a situation and so the paired t-test, being easier to compute, may be preferable.

Choosing an appropriate test may, therefore, depend on the amount of intracluster correlation present. If the intracluster correlation is small, the weighted tests are more powerful, since they take both the within-cluster and between-cluster variation into account. When the intracluster correlation is larger, the advantage in power of the weighted procedures becomes less since the between-cluster variation becomes more important and so an unweighted paired t-test is adequate.

The consequences of the effect of the variation between clusters and the treatment effect being combined in the model used for the weighted t-test was explained in a personal communication from Donner [126]. He explains that since $\hat{\rho}$ is spuriously large when a treatment effect exists, the estimate of $var(\hat{\theta}_i)$ using (149) will also be spuriously large for each stratum. Hence, a potential decrease in the power of the t-test would result. Donner states [126] that any choice of weights will lead to a valid test, but the efficiency is increased by estimating $var(\hat{\theta}_i)$ between, rather than within, strata when obtaining the variance of the estimate of the overall treatment effect. A reduction in power would occur if a within-stratum estimate of $var(\hat{\theta}_i)$ were used, due to the bias in the estimate of ρ in all the weighted procedures described where an estimate of the intracluster correlation is required. The use of a between-stratum estimate in these tests appears to be a means of correcting for this bias.

The generalisation of the paired t-test proposed by Rosner [81] is very similar to the random effects likelihood method in Section 2.2. In both these methods a consistent treatment effect is estimated separately from the between cluster variation, whereas in Donner's weighted methods the between cluster variation is confounded with the treatment effect. Since the estimate of ρ will be unbiased under the null hypothesis of no treatment effect, but under the alternative hypothesis it will be biased, the tests will still be valid although perhaps less powerful.

7.5 Results of a Comparison of the Tests

In order to compare the tests for binary outcomes described in Section 7.2 with each other and with the standard meta-analysis type tests, data from the British family heart study were used. A test derived from the quadratic approximation to the likelihood in Section 2.3.4, which is equivalent to Rosner's test for continuous outcome measures (Section 7.3.3), was also considered. This test is given by $\hat{\theta}_l / \sqrt{I_{11}^{*-1}(\hat{\theta}_l)}$ where $\hat{\theta}_l$ is the MLE of θ, which is, in most cases, approximately equal to $\hat{\theta}_l / \sqrt{var(\hat{\theta}_l)}$ since the covariance of $\hat{\theta}_l$ and $\hat{\sigma}^2_{Bl}$ will usually be negligible. This test statistic is then compared with a standard normal distribution, although it may be better to compare it with a $t_{(k-1)}$ distribution as suggested by Rosner [81]. The purpose of this practical example is to see whether the tests produce similar or different results and conclusions, and to consider the effect of differing amounts of heterogeneity. The unweighted procedures are expected to perform less well than the weighted procedures or the random effects procedure under conditions of moderate heterogeneity, while the standard Woolf and Mantel-Haenszel tests for $H_0 : \theta = 0$ may be expected to produce spuriously significant results.

The difference in prevalence of current cigarette smoking between the intervention and control group at the one year screening was the outcome chosen. Test results were obtained for both men and women using both the internal and external control group comparisons (Tables 63–66). The comparisons with the internal control group are not of a paired cluster randomised design, since within each practice individuals were randomised to one of two groups. However, there will still exist a component of between-town variation which may be estimated, unless the treatment effects in all towns are homogeneous.

For the comparison of the intervention group with the external control group, heterogeneity is present in the estimates for both sexes (Tables 63 and 64). The results

Table 63: Comparison of results of six different tests for the difference in the prevalence of cigarette smoking between the intervention and the external control group in men in the British family heart study

Test	Observed value of statistic	Distribution under H_0	p-value
Woolf	2.654	$N(0,1)$	0.008
Mantel-Haenszel	2.950	$N(0,1)$	0.003
Unweighted t	2.216	t_{k-1}	0.047
Weighted t	2.054	t_{k-1}	0.063
Weighted logistic	2.027	t_{k-1}	0.066
Permutation	2.051	$N(0,1)$	0.040
Rosner	2.054	$N(0,1)$	0.040

Test statistic for heterogeneity Q=25.224

Estimated between-study variance $\hat{\sigma}_B^2$=0.079

Estimated intracluster correlation $\hat{\rho}$=0.010

Table 64: Comparison of results of six different tests for the difference in the prevalence of cigarette smoking between the intervention and the external control group in women in the British family heart study

Test	Observed value of statistic	Distribution under H_0	p-value
Woolf	2.105	$N(0,1)$	0.035
Mantel-Haenszel	2.529	$N(0,1)$	0.011
Unweighted t	1.696	t_{k-1}	0.116
Weighted t	1.568	t_{k-1}	0.143
Weighted logistic	1.538	t_{k-1}	0.150
Permutation	1.585	$N(0,1)$	0.113
Rosner	1.618	$N(0,1)$	0.106

Test statistic for heterogeneity Q=22.114

Estimated between-study variance $\hat{\sigma}_B^2$=0.098

Estimated intracluster correlation $\hat{\rho}$=0.012

of the tests for women (Table 64) provide a clear example of what can happen when a positive intracluster correlation is present and a Mantel-Haenszel type procedure used to test the inappropriate null hypothesis $H_0 : \theta = 0$. Both the Woolf and the Mantel-Haenszel test, that is the tests which assume homogeneity of treatment effects across strata, give highly significant p-values, indicating evidence against the null hypothesis of no overall intervention effect. However, since there is heterogeneity present, the null hypothesis being investigated is $H_0 : \theta_i = 0$ for all i rather than $H_0 : \theta = 0$. The other five tests produce p-values greater than 0.1, providing much less evidence against the null hypothesis $H_0 : \theta = 0$. A similar effect, although not so clear because all tests produce apparent evidence against $H_0 : \theta = 0$, can be seen in the results of the tests for men (Table 63). All three t-tests, in this instance, produce similar results, with the permutation test and the Rosner type test appearing to be slightly more powerful in this example. Hence, it can be seen why the Mantel-Haenszel test and the Woolf test cannot be interpreted as tests of $H_0 : \theta = 0$ in cases where heterogeneity exists, whereas all the other tests appear to be adequate for such purposes.

There is no significant statistical heterogeneity when comparing the intervention group with the internal control group for either sex (Tables 65 and 66). Hence, even the Woolf and the Mantel-Haenszel test should be valid for the null hypothesis of $H_0 : \theta = 0$ in these cases and, furthermore, the Rosner type test based on the random effects model will be equivalent to the Woolf test. All tests provide evidence of a difference in the prevalence of cigarette smoking between the two groups for men and only the weighted t-test fails to detect a significant difference for women. The results indicate that the reported prevalence of cigarette smoking is lower in the intervention than in the control group. The result of the weighted t-test for women is rather odd as it is markedly different from the other results, including the other weighted procedure and no plausible explanation has been determined. There is only a small intracluster correlation and so the weighted procedures would, in fact, be expected to have greater power than the unweighted.

Table 65: Comparison of results of six different tests for the difference in the prevalence of cigarette smoking between the intervention and the internal control group in men in the British family heart study

Test	Observed value of statistic	Distribution under H_0	p-value
Woolf	3.021	$N(0,1)$	0.003
Mantel-Haenszel	3.057	$N(0,1)$	0.002
Unweighted t	4.473	t_{k-1}	0.001
Weighted t	4.738	t_{k-1}	0.000
Weighted logistic	4.418	t_{k-1}	0.001
Permutation	2.851	$N(0,1)$	0.004
Rosner	3.021	$N(0,1)$	0.003

Test statistic for heterogeneity Q=5.414

Estimated between-study variance $\hat{\sigma}_B^2$=0

Estimated intracluster correlation $\hat{\rho}$=0.001

Table 66: Comparison of results of six different tests for the difference in the prevalence of cigarette smoking between the intervention and the internal control group in women in the British family heart study

Test	Observed value of statistic	Distribution under H_0	p-value
Woolf	2.229	$N(0,1)$	0.026
Mantel-Haenszel	2.299	$N(0,1)$	0.026
Unweighted t	2.480	t_{k-1}	0.029
Weighted t	1.777	t_{k-1}	0.101
Weighted logistic	2.416	t_{k-1}	0.033
Permutation	2.099	$N(0,1)$	0.036
Rosner	2.229	$N(0,1)$	0.026

Test statistic for heterogeneity Q=10.482

Estimated between-study variance $\hat{\sigma}_B^2$=0

Estimated intracluster correlation $\hat{\rho}$=0.003

From these results, the unweighted paired t-test does appear to be robust and produces adequate results even when there is substantial intracluster correlation. It performs comparably with the weighted logistic t-test in the examples considered. The weighted t-test also gives similar results to the other two versions of the t-test apart from in one case in Table 66. The permutation test and the Rosner type test also produce results which are compatible with the t-tests. However, for this outcome, the intervention effect was small (Table 53) and so the estimate of the intracluster correlation will not be greatly inflated. In examples where there is a larger treatment effect, a fall in power of the weighted procedures would be expected, and hence the Rosner type test may be preferable in general.

7.6 Published Methods For Estimation in Paired Cluster Randomised Trials When the Outcome Variable is Dichotomous

When dealing with estimation of the overall treatment effect for dichotomous outcome variables, Donner and Klar [130] distinguish between two types of pair-matched cluster designs. Type (i) designs are those in which the cluster sizes may be relatively small, such as families, but the number of strata is reasonably large. Type (ii) designs are those in which the cluster sizes are fairly large, such as general practices, but the number of strata may be small. The reason for distinguishing between these two types of design is that they correspond to two different sets of asymptotic conditions. For type (i) designs, the asymptotic conditions assume that the stratum sizes are fixed but that the number of strata becomes large. In contrast, for type (ii) designs, they assume that the number of strata is fixed but that the sample sizes within each stratum become large. Since the type of design considered from a meta-analysis perspective will be of the type (ii) design, only the type (ii) designs will be considered here. This is the design where there are more problems in the analysis and where

modifications must be made to the standard methods.

Section 7.6.1 describes a modified Mantel-Haenszel estimator, while Section 7.6.2 describes a modification to the Woolf estimator.

7.6.1 Modified Mantel-Haenszel estimator

The standard Mantel-Haenszel estimator (Section 1.5.2) may be written in the following way

$$\hat{\theta}_{MH} = \frac{\sum_{i=1}^{k} w_i \hat{\theta}_i}{\sum_{i=1}^{k} w_i} \tag{180}$$

where $\hat{\theta}_i$ is the odds ratio in stratum i and $w_i = b_i c_i / N_i$ (notation in Table 2). The weight can be rewritten in terms of the numbers of individuals in stratum i, n_{i1} and n_{i2}, and the estimated proportions $\hat{P}_{i1} = a_i/n_{ij}$, $\hat{P}_{i2} = c_i/n_{i2}$ and $\hat{Q}_{ij} = 1 - \hat{P}_{ij}$ (j=1,2),

$$w_i = \frac{n_{i1} n_{i2}}{n_{i1} + n_{i2}} \hat{Q}_{i1} \hat{P}_{i2} \tag{181}$$

The standard Mantel-Haenszel estimator of the overall odds ratio is unbiased in large samples, even when clustering is present, but not in small samples [122]. However, the unmodified confidence limits, such as those given by Robins et al. [132], are not valid in type (ii) designs since the variance of the Mantel-Haenszel estimate requires that the number of strata be large (Robins et al. suggested as a practical criterion that the number of strata be at least 20).

Donner and Hauck [125] propose alternatives to these weights by considering the inflation in the variance of an individual cluster due to the clustering as a shrinking of the effective sample size. The actual sample size n_{ij} is replaced by the effective

sample size $n_{ij}^c = n_{ij}/[1+(n_{ij}-1)\rho]$ when calculating the weights (181). The effective sample size is such that when ρ=0, $n_{ij}^c = n_{ij}$ and when ρ=1, n_{ij}^c=1. For values of ρ between these two extremes, n_{ij}^c will take a value somewhere between one and the true sample size where the larger the intracluster correlation, the smaller the effective sample size. The intracluster correlation coefficient may be estimated as in Section 7.2.2 and hence the modified Mantel-Haenszel estimate of the overall odds ratio is given by

$$\hat{\theta}_{cMH} = \frac{\sum_{i=1}^{k} w_{im}\hat{\theta}_i}{\sum_{i=1}^{k} w_{im}} \tag{182}$$

where $w_{im} = n_{i1}^c n_{i2}^c \hat{Q}_{i1}\hat{P}_{i2}/(n_{i1}^c + n_{i2}^c)$. The weights w_{ic} are most suitable when θ=1 and depart from optimality as θ becomes large or small since $\hat{\rho}$ is only a consistent estimator of ρ if there is no treatment effect (Section 7.4). Hence, as in the case of hypothesis testing, when a treatment effect exists, $\hat{\rho}$ will be spuriously increased and so the effective sample sizes will be decreased and hence the weights w_{ic} will be larger than they should be. Hence, although the estimate will still be unbiased on average, it may be far from the true value in each specific case. There is no published method for obtaining the confidence interval of this modified Mantel-Haenszel estimate and so it is of very limited use.

7.6.2 Modified Woolf estimator

Donner and Klar [130] suggest as an alternative using the 'studentised Woolf method'. The estimate of the overall log odds ratio $\hat{\theta}$ is exactly that given in Section 1.5.1, where $var(\hat{\theta}_i) = (1/a_i)+(1/b_i)+(1/c_i)+(1/d_i)$, which may be rewritten as $[1/(n_{i1}\hat{P}_{i1}\hat{Q}_{i1})]+[1/(n_{i2}\hat{P}_{i2}\hat{Q}_{i2})]$. However, when calculating the variance of this estimate $\hat{\theta}$, instead of taking the variance of each individual study separately to be v_i=$1/w_i$ as in Section 1.5.1, a pooled between-stratum estimate of the variance of $\hat{\theta}_i$ is obtained. Hence

$$var(\hat{\theta}) = \frac{s^2 \sum_{i=1}^{k} w_i^2}{[\sum_{i=1}^{k} w_i]^2} \tag{183}$$

where $s^2 = \sum_{i=1}^{k} w_i(\hat{\theta}_i - \hat{\theta})^2 / \sum_{i=1}^{k} w_i$, thus taking the form common to all the weighted techniques proposed by Donner (Sections 7.2.2, 7.2.3 and 7.3.2). In the usual Woolf method (Section 1.5.1), when any clustering (or heterogeneity) is not taken in to account, the individual within stratum variances v_i are used as opposed to a weighted pooled estimate of the variance s^2. In this way the variance of the overall treatment effect reduces to $1/\sum_{i=1}^{k} w_i$.

The confidence intervals constructed using this standard variance may, however, be spuriously narrow when clustering is present and thus exaggerate the precision with which treatment effects are estimated. This leads to the modification of the estimator to produce the 'clustered Woolf estimator' which does allow for the effect of clustering both in the point estimate and confidence interval. The intracluster correlation coefficient ρ must again be estimated as shown in Section 7.2.2. Hence the 'clustered Woolf estimator' is still a weighted average of the individual study estimates, but with alternative weights,

$$\hat{\theta}_{cw} = \frac{\sum_{i=1}^{k} w_{ic} \hat{\theta}_i}{\sum_{i=1}^{k} w_{ic}} \tag{184}$$

where $\hat{\theta}_i$ is the log odds ratio for stratum i and $w_{ic} = [1/(n_{i1}^c \hat{P}_{i1} \hat{Q}_{i1}) + 1/(n_{i2}^c \hat{P}_{i2} \hat{Q}_{i2})]^{-1}$, that is the reciprocal of the variance of $\hat{\theta}_i$ with n_{ij} replaced by n_{ij}^c (Section 7.6.1). The estimate of the variance of $\hat{\theta}_{cw}$ can then be used to obtain the confidence limits for this estimate and is given by

$$var(\hat{\theta}_{cw}) = \frac{s_c^2 \sum_{i=1}^{k} w_{ic}^2}{[\sum_{i=1}^{k} w_{ic}]^2} \tag{185}$$

where $s_c^2 = \sum_{i=1}^{k} w_{ic}(\hat{\theta}_i - \hat{\theta}_{cw})^2 / \sum_{i=1}^{k} w_{ic}$. Again a between-stratum estimate of $var(\hat{\theta}_i)$ is used rather than estimating it for each stratum separately. A simulation study carried out by Donner and Hauck [125] with binary data suggests that the Woolf method can be recommended for designs having six or more strata and at least 40 subjects per cluster. However, in general, the modified estimator has been found to be more precise, that is to have a smaller mean square error, than the unmodified estimator in the type (ii) paired cluster designs [125].

7.7 Published Methods For Estimation in Paired Cluster Randomised Trials When the Outcome Measure is Continuous

Donner and Klar [130] propose the use of either an approach based on an unweighted paired t-test or one based on a weighted paired t-test. In the situation of a continuous outcome variable, the measure of treatment effect is the difference in means between the two treatment groups. Using the notation and the ideas presented in Section 7.7, estimates, variances and, therefore, confidence limits may be obtained. For the unweighted paired t-test (Section 7.3.1), the overall estimate of treatment effect is obviously given by $\bar{d} = \sum_{i=1}^{k} \hat{\theta}_i / k$ and then the variance of $\bar{d}$ is given by

$$\frac{s^2}{k} \tag{186}$$

Furthermore, the weighted estimate of the overall difference is $\bar{d}_w = \sum_{i=1}^{k} w_{iw}\hat{\theta}_i / \sum_{i=1}^{k} w_{iw}$ (Section 7.3.2) and the corresponding variance is given by (174)

As in the case of testing and estimation for dichotomous outcomes in methods where the intracluster correlation is used, both methods outlined in this section are affected

by the bias in the estimate of ρ when a treatment effect exists.

7.8 Results

The example outcome variable from the British family heart study of prevalence of current cigarette smoking which was used when considering testing (Section 7.5) will also be used to compare the different methods of estimation in order to see whether the clustering must be taken into account. The three methods (modified Mantel-Haenszel, unmodified and modified Woolf) described in Section 7.6 were considered and compared with the Woolf method where $var(\hat{\theta}_i) = 1/w_i$, instead of s^2, is used to obtain $var(\hat{\theta}_i)$ (Section 1.5.1) and the random effects meta-analysis method using the D&L moment estimator of σ_B^2. A 'clustered Woolf method' was also considered taking $var(\hat{\theta}_i)$ to be $1/w_{ic}$, rather than s_c^2, thus implying that the variance of the estimate of the overall treatment effect would be $1/\sum_{i=1}^{k} w_{ic}$, since this would appear to be the natural estimator for the variance of the clustered estimator.

The modified Mantel-Haenszel estimates are fairly similar to the unmodified ones (Table 67), but the real difference would be in the precision of the two estimates. However, since there is no apparent method of obtaining the confidence interval for the modified estimate, then a real comparison is not possible. Furthermore, this lack of a measure of precision on the modified estimate severely limits its usefulness.

Since there is very little intracluster correlation for the internal control group comparison in men, all methods, both Mantel-Haenszel type and Woolf type, produce very similar estimates as would be expected (Tables 67 and 68). However, the two Donner and Klar Woolf methods, that is those where $var(\hat{\theta}_i)$ is estimated by a pooled between-stratum variance s^2, produce much smaller variances of the overall estimate than the methods using individually estimated variances (Table 68). This behaviour is also evident in the internal control group comparison for the women, although

Table 67: Comparison of two Mantel-Haenszel type estimates of the overall odds ratio comparing the prevalence of cigarette smoking in the intervention group and the control groups in the British family heart study

Mantel-Haenszel estimator	Estimate of overall odds ratio (variance)			
	Men		Women	
	Internal	External	Internal	External
Unmodified	0.784	0.807	0.793	0.793
	(0.0068)	(0.0058)	(0.0109)	(0.0091)
Modified (*)	0.784	0.787	0.784	0.797

* No method for obtaining the variance

the difference in the variances between the two types of method are less noticeable (Table 68). Hence, in the examples where there is no heterogeneity present, the variances for the clustered estimate obtained using the between-stratum estimate of $var(\hat{\theta}_i)$ are too small, while those obtained using within-stratum estimates are too large. Therefore, when there is no significant clustering, a standard Woolf method is preferable to a clustered method, since the associated variance will be more reliable. The clustered variance proposed in the literature will produce a variance which is too small, thus providing stronger evidence of a treatment effect than actually exists.

It is in the case of the two external control group comparisons that the modified estimates are really required since substantial clustering is present for both men and women. For men, the clustered estimate agrees well with the random effects estimate, with the meta-analysis estimator producing a slightly smaller overall log odds ratio (Table 68). All variances are similar with the exception of the standard Woolf method using individually estimated weights for each strata, which is too small. This is to be

Table 68: Comparison of Woolf type estimates of the overall odds ratio comparing the prevalence of cigarette smoking in the intervention group and the control groups, together with variances, in the British family heart study

Estimator	Estimate of overall odds ratio (variance)			
	Men		Women	
	Internal	External	Internal	External
standard $var(\hat{\theta}_i) = s^2$	-0.242	-0.197	-0.228	-0.198
	(0.0030)	(0.0119)	(0.0100)	(0.0179)
standard $var(\hat{\theta}_i) = 1/w_i$	-0.242	-0.197	-0.228	-0.198
	(0.0064)	(0.0055)	(0.0104)	(0.0088)
clustered $var(\hat{\theta}_i) = s_c^2$	-0.242	-0.226	-0.240	-0.203
	(0.0030)	(0.0120)	(0.0099)	(0.0164)
clustered $var(\hat{\theta}_i) = 1/w_{ic}$	-0.242	-0.226	-0.240	-0.203
	(0.0072)	(0.0152)	(0.0134)	(0.0220)
random effects	-0.242	-0.228	-0.227	-0.213
	(0.0064)	(0.0119)	(0.0104)	(0.0171)

expected, since this is a fixed effect model that does not take into account the extra random variation. It is not clear, however, why the variance for the standard fixed effect estimate derived using s^2 is so much larger than that using $1/w_i$, as this method does not take account of clustering either. This point is raised in the discussion section (Section 7.9).

When using $var(\hat{\theta}_i) = 1/w_{ic}$, the variance of the estimate of overall treatment effect is spuriously increased in cases where heterogeneity is present due to the previously discussed bias in the estimation of ρ (Table 68). Although estimating $var(\hat{\theta}_i)$ using the between-stratum variance would not appear to be based on sound statistical theory since it assumes that the variance for each strata estimate is the same, it does appear to produce more reliable estimates for the variance of $\hat{\theta}_{cw}$ by causing a decrease in $var(\hat{\theta}_{cw})$. The variances obtained in this way are close to those obtained from the random effects meta-analysis when heterogeneity is present. Therefore, it appears to correct approximately for the bias in the estimate of ρ, although it is not obvious how or why. This is investigated further in Section 7.9. Hence, a random effects meta-analysis method would appear preferable when heterogeneity is present.

7.9 Discussion

Due to the fact that when a large treatment effect exists, the estimate of ρ is biased, then the estimate of each weight w_{ic} will also be biased downwards. The examples in Section 7.8 show how the results obtained from the paired cluster randomised methods could be misleading, as the estimates and variances from the cluster randomised methods do not always agree with those from the random effects methods (Table 67). When a treatment effect is present, but there is no evidence of heterogeneity, $\hat{\rho}$ will be biased thus introducing extra variation which does not exist. Hence, it causes the variance associated with the estimate of overall treatment effect, as calculated by Donner and Klar [130], to be too small. This additional variance will also mean

that although the estimator of the overall treatment effect is still unbiased, the estimate obtained in practice may be a long way from the true value. When there is heterogeneity present, the variance of the estimate of the overall treatment effect may still be incorrect as the extra variation will be overestimated when a treatment effect exists. However, by using a between-stratum estimate of $var(\hat{\theta}_i)$ a reasonable estimate of the variance of $\hat{\theta}_{cw}$ appears to be obtained.

In the example from the British family heart study, the treatment effect is fairly small and hence the estimate of ρ will not be greatly biased. A hypothetical trial was therefore created in order to illustrate more clearly the failings of the paired cluster randomised methods. In this example the treatment effects, in terms of a log odds ratio, were large (varying between -0.8 and -0.4) in each of the 13 strata, but were also homogeneous. The Q statistic for heterogeneity was only 3.858 and so the between-stratum variance was set to 0.

However, due to the large treatment effect, the estimate of $\hat{\sigma}^2_{BC}$ was 0.0041 and so the estimate of the intracluster correlation was greater than 0. The fixed effect (or equivalently in this example the random effects) meta-analysis estimate of the overall log odds ratio was larger than the clustered Woolf estimate (Table 70). The reduction in the overall odds ratio is due to the different allocation of weight between the strata in the standard fixed effect meta-analysis method and the clustered Woolf method (Table 71). The clustered Woolf method, because of the extra spurious variation it introduces, gives more weight to the smaller imprecise stratum estimates, all of which happen to be smaller than -0.5, and so the overall estimate is decreased. This example backs up the findings from the British family heart study example (Section 7.8) regarding $var(\hat{\theta}_{cw})$ as defined by Donner and Klar [130], since $var(\hat{\theta}_{cw})$ obtained using a between-stratum estimate of $var(\hat{\theta}_i)$ produces a value which is smaller than that obtained using the standard fixed effect meta-analysis method. Hence, when there is homogeneous data, using the between-stratum estimate of $var(\hat{\theta}_i)$ produces an es-

Table 69: Data for a hypothetical example with 13 centres where the treatment effect is large but where there is no heterogeneity

Centre number	Events/total number of patients		Odds Ratio
	Treated	Control	
1	25/100(25%)	50/100(50%)	0.500
2	2/40(5%)	5/50(10%)	0.500
3	60/250(24%)	80/200(40%)	0.600
4	20/200(10%)	45/200(22.5%)	0.440
5	6/50(12%)	6/25(24%)	0.500
6	25/150(16.7%)	40/150(26.7%)	0.625
7	55/500(11%)	100/500(20%)	0.550
8	75/300(25%)	150/400(37.5%)	0.667
9	2/50(4%)	4/50(8%)	0.500
10	100/600(16.7%)	150/600(25%)	0.667
11	12/100(12%)	25/100(25%)	0.480
12	14/200(7%)	30/200(15%)	0.467
13	75/800(9.4%)	75/500(15%)	0.625

timate of $var(\hat{\theta}_{cw})$ which is too small. However, it is a better estimate of $var(\hat{\theta}_{cw})$ than that which would be obtained if the within-strata variance estimates were used leading to $var(\hat{\theta}_{cw})$ being equal to $1/\sum_{i=1}^{k} w_{ic}$. This estimate would be too large, and in this particular example would be far too large, taking the value 0.74 (Table 70).

Table 70: Comparison of Woolf type estimates of the overall odds ratio in the hypothetical example and the diuretics trials example

Estimator	Estimate of overall log odds ratio ($\hat{\theta}$) (variance)	
	Hypothetical	Diuretics trials
standard $var(\hat{\theta}_i) = s^2$	-0.52	-0.40
	(0.036)	(0.072)
standard $var(\hat{\theta}_i) = 1/w_i$	-0.52	-0.40
	(0.004)	(0.008)
clustered $var(\hat{\theta}_i) = s_c^2$	-0.60	-0.57
	(0.002)	(0.082)
clustered $var(\hat{\theta}_i) = 1/w_{ic}$	-0.60	-0.57
	(0.74)	(1.308)
random effects	-0.52	-0.51
	(0.004)	(0.042)

standard estimate=$\sum_{i=1}^{k} w_i\hat{\theta}_i/\sum_{i=1}^{k} w_i$
clustered estimate=$\sum_{i=1}^{k} w_{ic}\hat{\theta}_i/\sum_{i=1}^{k} w_{ic}$

The diuretics trial data (Section 1.3.1) (even though it is not of a paired cluster randomised design) may be used as a second example, by considering each group to be a cluster, to illustrate what occurs when there is both a large treatment effect and a large amount of heterogeneity in the data. In this case, therefore, a trial represents a stratum and a treatment group a cluster. Due to the large amount of heterogeneity

Table 71: A comparison of the percentage weight allocated to each centre in the random effects meta-analysis method and the paired cluster randomised method for the hypothetical example and the diuretics trials example

Centre number	Percentage of total weight $(w_i/\sum_{i=1}^{k} w_i) \times 100$			
	Hypothetical example		Diuretics trials	
	Random effects estimate (*)	Cluster estimate	Random effects estimate	Cluster estimate
1	5.3	12.0	10.7	9.7
2	0.6	3.7	11.9	7.3
3	11.0	11.4	10.2	21.9
4	5.1	6.9	7.9	17.7
5	1.1	7.6	12.1	1.9
6	5.4	9.1	17.0	10.4
7	13.1	7.0	11.8	3.3
8	16.3	11.4	4.5	2.9
9	0.5	3.1	13.9	24.9
10	21.0	9.0		
11	2.9	7.7		
12	3.7	5.1		
13	14.0	6.0		

* Since the estimate of the between-study variance is 0 the random effects model is the same as the fixed effect model in this example

both the meta-analysis estimate of the between-stratum variance and the intracluster correlation are greater than zero. The estimates of the overall log odds ratio and their variances are different using the different methods. As for the previous example it may be seen that the weights are allocated in different ways in the two different methods (Table 71) with the clustered method appearing to allocate higher weights to the trials with the largest event rates, that is the largest $\hat{P}_{ij}$, rather than those with the greatest precision. The clustered method produces a point estimate which is between the fixed effect and the random effects estimates. The variance of $\hat{\theta}_{cw}$ calculated as proposed by Donner and Klar [130] is considerably larger than the variance obtained from the random effects model (Table 70). However, if $1/\sum_{i=1}^{k} w_{ic}$ were to be used instead, the variance would be far too large because of the large inflation in the value of $\hat{\rho}$. This example shows that in the presence of a large amount of heterogeneity and a large treatment effect the clustered Woolf method, even with the adjustment to the $var(\hat{\theta}_{cw})$ of using s_c^2, may produce a variance for the overall estimate which is too large (Table 70).

This effect of an underestimation of the variance when there is no heterogeneity but an overestimation when there is heterogeneity is due to the use of a between-stratum estimate of $var(\hat{\theta}_i)$. This appears to be an approximate way of correcting for the bias in the estimate of ρ, since without this adjustment the variances using $1/\sum_{i=1}^{k} w_{ic}$ would be far too large. This correction may be adequate when there is a small treatment effect, as in the example from the British family heart study (Section 6.3.3), but is obviously poor when a large treatment effect exists, irrespective of the amount of heterogeneity. The reason for the differences in the variances of $\hat{\theta}_{cw}$ using s^2 compared to the individual v_i, is the difference in the values that an average within-stratum variance and a between-stratum variance estimate takes. When there is no heterogeneity a between-stratum estimate will be smaller than the average within-stratum estimate. However, when there is heterogeneity, the between-stratum estimate will tend to be large and will produce an estimate which is

larger than the average within-stratum estimate (Table 72). It should be noted that for the hypothetical example the mean variance is simply $\overline{v}_i$ as σ_B^2=0.

Table 72: Comparison of the mean within-cluster estimate and the pooled estimate of $var(\hat{\theta}_i)$

Example	Average within-study variance $(\overline{v_i + \sigma_B^2})$	Pooled variance (s_c^2)
Hypothetical	0.19793	0.01912
Diuretics trials	0.42505	0.48354

7.10 Conclusion

For estimation of an overall treatment effect, meta-analysis methods applied to paired cluster randomised designs have a clear advantage over the methods supposedly designed specifically for such trials discussed in this chapter. It has been shown that the estimation methods relying on the estimation of an intracluster correlation are biased when a large treatment effect exists. The adjustment to the variance which is apparently made to correct for the bias in $\hat{\rho}$ is not founded on solid statistical theory and, is in any case, unreliable when the treatment effect is large and there is considerable heterogeneity. Hence, as was shown in the two examples of Section 7.9, conclusions drawn from the analysis can be misleading. Under conditions of homogeneity, the variances used by Donner and Klar [130] will produce confidence intervals which are too narrow and hence the possibility of obtaining a spurious significant result exists. On the other hand, if heterogeneity is present, then the increased variance may lead to a conservative interpretation of the findings. The random effects meta-analysis method does not have the problem that the random variation is influenced by the size of the treatment effect, because the treatment effect and the between-study variance are estimated separately. Hence, meta-analysis methods are an improvement on

the standard methods in the analysis of paired cluster randomised trials.

With respect to testing the null hypothesis $H_0 : \theta = 0$, then depending on the amount of heterogeneity present, different tests are more suitable than others under different conditions. If homogeneity can be assumed, then the Mantel-Haenszel test is optimal [38], while at the other extreme, if there is considerably larger between-cluster than within-cluster variation, a simple unweighted t-test is adequate. Due to the small amount of variation within each cluster such a case is effectively equivalent to obtaining a single observation from each cluster. Between these two contrasting cases, where there is some heterogeneity, a weighted procedure is perhaps most suitable, since it takes account of both types of variation, which is what is required when neither dominates. However, due to the bias in the estimate of ρ and the inadequacy of the correction to the variance of the overall treatment effect using a between-cluster estimate of $var(\hat{\theta}_i)$, the tests will lack power when heterogeneity exists. Hence, a test based on the random effects model of Section 2.2), and similar to that proposed by Rosner, is probably more reliable for all situations and will reduce to the standard fixed effect test under homogeneity.

Hence, overall for both testing and estimation, the use of meta-analysis techniques in the analysis of paired cluster randomised trials provides a clear improvement over the method currently proposed in the literature. There is a fundamental flaw in the procedure for the estimation of the intracluster correlation. Although such methods may be adequate in certain cases, they are in general unreliable and may produce misleading results in many circumstances.

8 Conclusions

The research outlined in this thesis has covered various different statistical issues relating to meta-analysis in medical research, and has also shown how a meta-analysis approach is useful in the analysis of multicentre trials and paired cluster randomised trials. This final chapter contains a summary of the conclusions from each of the previous chapters and brings some of the ideas together in a more general discussion and also highlights the practical implications.

In Chapter 1 the two standard meta-analysis models were introduced, that is the fixed effect and the random effects models, and the drawbacks of each of these approaches was highlighted. This served as a starting point for the development of much of the research in the following four chapters. Chapter 2 focused on the random effects model and introduced a likelihood approach to meta-analysis based on the marginal likelihood of each trial. It is concluded that such an approach may be required in practice, particularly in situations where the between-study variance is imprecisely estimated and if changes in the between-study variance have an effect on the estimate of the overall treatment effect. Sensitivity plots, described in Section 2.1, are useful not only in their own right for the purpose of investigating the robustness of the conclusions drawn from the standard meta-analysis models to changes in σ_B^2 but also for investigating whether a likelihood method may be required. It is necessary to carry out such checks as the use of the likelihood model may lead to a more conservative interpretation of the effectiveness of the treatment due to a wider confidence interval for θ being obtained. Unlike the standard random effects model, this likelihood model also allows a confidence interval to be obtained for the between-study variance so that the precision of this estimate may be summarised directly. In practice, $\hat{\sigma}_B^2$ was often found to be imprecise, thus reinforcing the need for a sensitivity analysis.

This marginal likelihood approach was also compared with the Mantel-Haenszel-

type likelihood approach of van Houwelingen et al. [45] which is based on the exact distribution of each trial's 2×2 contingency table (Section 2.4). Although this full likelihood model is a better representation of the data than the marginal likelihood, and furthermore does not make the assumption about weights being known, the results from the two methods have, for the examples considered, been found to be comparable. However, the Mantel-Haenszel-type method does have the advantage of being able to deal with zero event rates and small frequencies (Section 2.5) unlike the marginal likelihood method where empirical logits would be required in the presence of small frequencies. On the other hand, the marginal likelihood method is more flexible in that it can be used to analyse continuous as well as binary outcome measures. Overall, in most situations where either method could be used, the two likelihood procedures will produce very similar results. Hence, in general the choice between the methods may depend more on the practical concerns regarding the implementation of the procedures (Section 2.8).

Empirical Bayes or fully Bayesian approaches offer alternative ways for considering the meta-analysis problem. Empirical Bayes estimates may be more useful in practical terms than an estimate of the overall treatment effect when a range of estimates that could be obtained is required to make clinical judgements about the appropriateness of treatment. A fully Bayesian approach has the advantage that it overcomes the need for having to make the controversial assumption that the trials included in a meta-analysis are a random sample from a large population of trials. A Bayesian approach, however, is considerably more computer intensive than either the standard fixed effect or random effects approaches, or even the likelihood methods. The Bayesian meta-analysis literature does provide discussion on ways of looking at the robustness of the conclusions from the meta-analysis. Indeed, sensitivity analyses and the checking of assumptions have been discussed more in the Bayesian than the classical framework, but work in this thesis has shown ways in which the conclusions may be checked for robustness and the modelling assumptions in a classical statistical

approach.

Part of this work was described in Chapter 3 which set out to provide ways in which to check the assumption of normality, an assumption necessary to produce confidence intervals for both the fixed effect and random effects estimate of the overall treatment effect. The use of q-q plots of the q_i and of the q_i^* components of Q were proposed to check these assumptions. Furthermore, a test of the null hypothesis $q_i \sim N(\mu, 1)$ using an Anderson-Darling statistic was found to be the best test of the normality assumptions. It can test the gradient of the plot as well as the linearity, thus enabling it to distinguish between a fixed effect and a random effects model on a fixed effect plot as well as being able to detect non-normality. In general, consideration of both plots and the results of the Anderson-Darling test for normality is adequate to establish whether the data follow either of the standard normally distributed models. However, if it has been established that a set of data does not follow either of the standard normally distributed models, the question of how to proceed then arises. Further research, which could build on that presented in Chapter 3, is required to investigate whether the invalidity of the normal assumptions can affect the results of a meta-analysis. This could take the form of an investigation into the robustness of the results to deviations from the standard models using simulation methods such as those used in Chapter 3. Alternatively, the results from a non-parametric approach, such as that of van Houwelingen et al. [45] could be compared to those from standard methods under various alternative models. Further investigation into the use of alternative distributions for the random effects is also required.

Both plots and tests were found to be of limited use when the number of trials in the meta-analysis was small, for example when k=9 in the diuretics trials data. The q-q plots may still be useful, however, in identifying sources of heterogeneity which can then be investigated, although the same information may also be derived from a Galbraith plot [66]. In practice, a meta-analysis should always be accompanied

by investigations of the modelling assumptions and sources of heterogeneity, as well as by sensitivity analyses. Once heterogeneity has been identified work should be undertaken to try to explain the reasons behind the observed variation in terms of, for example, trial characteristics, population differences and geographic differences. By identifying trial-specific characteristics which explain the variation in treatment effect the analysis can be reduced to an analysis of homogeneous sets of data. Although the explanation of heterogeneity may be a commendable aim in principle, however, it can be difficult to accomplish satisfactorily in practice as exemplified by the analysis of the British family heart study in Chapter 6. Furthermore, it should be borne in mind that such investigations are always post-hoc and any resulting explanations will usually have been motivated by the observation of the data themselves. Hence, caution should be expressed in any conclusions drawn, particularly when there may be several alternative but equally feasible explanations. The findings from these investigations may be useful for motivating future research by identifying subgroups of patients for whom the treatment may be more, or less, effective. Ideally factors which are possible causes of heterogeneity should be identified prior to the actual analysis and prior to observation of the data, although this is not always realistic in practice. This is in line with the suggestion that a protocol should be drawn up before a meta-analysis is carried out [6], in a similar way to a protocol for a clinical trial, outlining how the meta-analysis is to proceed.

The choice between a fixed effect or a random effects approach, is often based on the result of the test, using Q, of heterogeneity of the $\hat{\theta}_i$. However, it was shown in Chapter 4 how the power of the test may be particularly low in certain conditions. It was also shown that a test proposed as an improvement to Q [110] was found to provide no substantial increase in power. Furthermore, the null distribution of this alternative statistic remained unclear, and hence the test using Q was still regarded as preferable. The power of the test for heterogeneity Q was found to be poor when the total amount of information available in the meta-analysis $\sum_{i=1}^{k} w_i$ was small, due

to either k being small or the trial estimates being imprecise, that is having large v_i. However, situations where extra care in the interpretation of the test result is required *are*where there is an uneven distribution of the weight between trials. This is because the power of the test was found to be particularly poor, for a given total amount of information and a given amount of between-study variation, when one trial took most of the weight and the other trials all produced very imprecise estimates of treatment effect. Hence, investigation of heterogeneity should be considered, particularly in such cases as that outlined above, even in the presence of a non-significant result for the overall test of heterogeneity.

Chapter 5 served to illustrate, in the context of continuous outcome measures, that problems may be caused by the assumption made in both the standard meta-analysis methods that the weights are known as opposed to estimated. It was concluded that the variances of both fixed effect and random effects estimates of overall treatment effect are incorrect when w_i are estimated. This may lead to a false certainty in the conclusions drawn from the fixed effect method in that the confidence interval obtained will be spuriously narrow. The confidence interval for the random effects model tends to be spuriously large due to the overestimation of the between-study variance. In practice, the effect will often be negligible, although caution is required when all or some of the n_i are small. It may be better to use the alternative methods for calculating the results which are based on the assumption that $\hat{w}_i$ is known and is equal to $f_i w_i$, that is the expectation of $\hat{w}_i$ where f_i is a correction factor for the estimation equal to $(n_i - 1)/(n_i - 3)$, rather than equal to simply w_i. In certain circumstances, that is when n_i are reasonably large the adjusted approximate results presented in Chapter 5 offer improvements in the performance. Further work is, however, required to refine the estimates to allow exactly rather than approximately for the estimation of w_i, and also to investigate the effect of the estimation of w_i on meta-analyses with binary outcome measures. However, this problem may perhaps more usefully be investigated further by a comparison of the standard

meta-analysis results with those from the Mantel-Haenszel-type likelihood model of van Houwelingen et al. [45].

It has been shown in this thesis how meta-analysis methods may be used to analyse single clinical trials with multiple centres, thus allowing for the possibility of variations in treatment effect across centres. Individually randomised trials may obviously be analysed in such a way by considering each centre as a 'trial', but so may paired cluster randomised trials as shown in Chapter 6 in relation to the British family heart study. It has been shown how difficult it may be to understand the reasons for any heterogeneity observed, particularly in multicentre trials where any practical differences between centres are less obvious due to the fact that all centres follow the same protocol. Hence, it may be concluded that in certain cases, such as the British family heart study, the variation can reasonably be considered as random, and a random effects model will provide the most satisfactory approach to analysis.

Analysing such trials as the British family heart study also raises the problem of multiple outcome measures. This issue was briefly considered in relation to the analysis of the British family heart study, in Section 6.5. A directional test of the null hypothesis that each outcome in each trial is zero may be useful in certain circumstances (Section 6.5.3) as it produces a test of the overall impact of the treatment. A generalised least squares model based on effect sizes (Section 6.5.2) may also be of some use in homogeneous sets of data, particularly where different outcomes are measured in different studies. However, further research is required on this topic to provide satisfactory solutions by extending the procedure to cope with random effects and different measures of treatment effect.

A random effects meta-analysis approach to the analysis of paired cluster randomised trials is certainly to be recommended over other methods which have been specifically designed for the analysis of such trials (Chapter 7). Approaches to both testing (Sections 7.2 and 7.3) and estimation (Sections 7.6 and 7.7) using the concept

of the intracluster correlation ρ (Section 7.1) were found to be biased in circumstances where a large treatment effect exists. This bias is due to the estimate of ρ being confounded by the estimate of the treatment effect, and hence such procedures may produce very misleading results. Although the corresponding tests are still valid under the null hypothesis of no treatment effect, they will be of low power. In terms of estimation, the confidence intervals for the estimate of the overall treatment effect can be incorrect (Section 7.6). In homogeneous sets of data, heterogeneity may be introduced due to the bias in $\hat{\rho}$ and confidence intervals which are too wide may be obtained, thus leading to the increased possibility of a misleading nonsignificant result. The opposite effect occurs when heterogeneity is present in that the confidence interval obtained is too narrow due to an overcorrection in the calculation of the variance of the estimate of the overall treatment effect for the bias in $\hat{\rho}$. The random effects meta-analysis approach has no such problems as it estimates the between-centre component of variation separately from the treatment effect. The generalisation of the paired t-test proposed by Rosner (Section 7.3.3) is more reliable since it is not based on the intracluster correlation. In fact, the model on which the test is based is that of the marginal likelihood of Section 2.2, where the quadratic approximation (Section 2.3.4) is made.

Meta-analysis methodology is useful for both combining information from different centres in a single trial as well as from different trials in a true meta-analysis. It is likely that the need for meta-analyses will grow in the future, partly because smaller treatment benefits will require detection, and also because carrying out single trials which are large and powerful enough may not be practical. Development of computer software for meta-analyses should not obscure the need for careful consideration in every specific analysis as to which trials to include, and also whether the trials can meaningfully be combined to produce an overall estimated treatment effect. Furthermore, an investigation of heterogeneity should always be included in any analysis. If no feasible explanation of heterogeneity is possible then a random effects analysis may

be the best practical alternative when heterogeneity exists, certainly more appropriate than a fixed effect estimate, since it does produce more appropriate confidence intervals, providing the distributional assumptions are valid. However, standard results are best accompanied by a sensitivity analysis indicating how the conclusions change as the between-study variance changes. Hence, meta-analyses cannot simply be reduced to the following of a set formula, whereby an overall estimate is obtained without cautionary investigations and discussion relevant to the individual case. Both the fixed effect and the random effects models are not ideal, but they will, in general, produce reliable results provided they are used in conjunction with appropriate supporting investigations and interpreted with the required degree of caution.

References

[1] Glass, G.V. 'Primary, secondary and meta-analysis of research'. *Educ Res*, **5**:3–8, (1976).

[2] Felson, D.T. 'Bias in meta-analytic research'. *J Clin Epidemiol*, **45**:885–892, (1992).

[3] Chalmers, T.C., Matta, R.J., and Smith Jr., H. 'Evidence favouring the use of anticoagulants in the hospital phase of acute myocardial infarction'. *N Engl J Med*, **297**:1091–1096, (1977).

[4] Whitehead, A. and Whitehead, J. 'A general parametric approach to the meta-analysis of randomized clinical trials'. *Stat Med*, **10**:1665–1677, (1991).

[5] Peto, R. 'Why do we need systematic overviews of randomized trials?'. *Stat Med*, **6**:233–240, (1987).

[6] Boissel, J-P., Blanchard, J., Panak, E., Peyrieux, J-C., and Sacks, H. 'Considerations for the meta-analysis of randomized clinical trials'. *Contr Clin Trials*, **10**:254–281, (1989).

[7] DeMets, D.L. 'Methods for combining randomized clinical trials: strengths and limitations'. *Stat Med*, **6**:341–348, (1987).

[8] Meinert, C.L. 'Meta-analysis: science or religion?'. *Contr Clin Trials*, **10**:257S–263S, (1989).

[9] Mulrow, C.D. 'Rationale for systematic reviews'. *Brit Med J*, **309**:597–599, (1994).

[10] L'Abbé, K.A., Detsky, A.S., and O'Rourke, K. 'Meta-analysis in clinical research'. *Ann Intern Med*, **107**:224–233, (1987).

[11] Chalmers, T.C., Berrier, J., Sacks, H.S., Levin, H., Reitman, D., and Nagalingam, R. 'Meta-analysis of clinical trials as a scientific discipline II: Replicate variability and comparison of studies that agree and disagree'. *Stat Med*, **6**:733–744, (1987).

[12] Thompson, S.G. 'Controversies in meta-analysis: The case of trials of serum cholesterol reduction'. *Stat Meth Med Res*, **2**:173–192, (1993).

[13] Naylor, C.D. 'Two cheers for meta-analysis: Problems and opportunities in aggregating results of clinical trials'. *Can Med Ass J*, **138**:891–895, (1988).

[14] O'Rourke, K. and Detsky, A.S. 'Meta-analysis in medical research: Strong encouragement for higher quality in individual research efforts'. *J Clin Epidemiol*, **10**:1021–1024, (1989).

[15] Rosenthal, R. 'The "file draw" problem and tolerance for null results'. *Psychol Bull*, **86**:638–641, (1979).

[16] Stewart, L.A. and Parmar, M.K.B. 'Meta-analysis of the literature or of individual patient data: Is there a difference?'. *Lancet*, **341**:418–422, (1993).

[17] Detsky, A.S., Naylor, C.D., O'Rourke, K., McGeer, A.J., and L'Abbé, K.A. 'Incorporating variations in the quality of individual randomized trials into meta-analysis'. *J Clin Epidemiol*, **45**:255–265, (1992).

[18] Emerson, J.D., Burdick, E., Hoaglin, D.C., Mosteller, F., and Chalmers, T.C. 'An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials'. *Contr Clin Trials*, **11**:339–352, (1990).

[19] Schulz, K.F., Chalmers, I., Hayes, R.J., and Altman, D.G. 'Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials'. *JAMA*, **273**:408–412, (1995).

[20] Klein, S., Simes, J., and Blackburn, G.L. 'Total parenteral nutrition and cancer clinical trials'. *Cancer*, **58**:1378–1386, (1986).

[21] Greenland, S. 'Invited commentary: A critical look at some popular meta-analytic methods'. *Am J Epidemiol*, **140**:290–296, (1994).

[22] Thacker, S.B. 'Meta-analysis. A quantitative approach to research integration'. *JAMA*, **259**:1685–1689, (1988).

[23] Lau, J., Antman, E.M., Jimenez-Silva, J., Kupelnick, B., Mosteller, F., and Chalmers, T.C. 'Cumulative meta-analysis of therapeutic trials for myocardial infarction'. *New Engl J Med*, **327**:248–254, (1992).

[24] Chalmers, I., Hetherington, J., Newdick, M., Mutch, L., Grant, A., Enkin, M., Enkin, E., and Dickerson , K. 'The Oxford Database of Perinatal Trials: developing a register of published reports of controlled trials '. *Contr Clin Trials*, **7**:306–324, (1986).

[25] Olkin, I. 'Invited commentary: Re: A critical look at some popular meta-analytic methods'. *Am J Epidemiol*, **140**:297–299, (1994).

[26] Sackett, D., editor. *Cochrane Collaboration Handbook*. Cochrane Collaboration, Oxford, 1994.

[27] Thompson, S.G. 'Why sources of heterogeneity in meta-analysis should be investigated'. *Brit Med J*, :1351–1355, (1994).

[28] Greenland, S. 'Quantitative methods in the review of epidemiologic literature'. *Epidemiol Rev*, **9**:1–30, (1987).

[29] Jones, D.R. 'Meta-analysis of observational epidemiological studies: A review'. *J Royal Soc Med*, **85**:165–168, (1992).

[30] Cochran, W.G. 'The combination of estimates from different experiments'. *Biometrics*, **10**:101–129, (1954).

[31] Stijnen, T. and van Houwelingen, H.C. 'Empirical Bayes methods in clinical trials meta-analysis'. *Biometrical Journal*, **32**:335–346, (1990).

[32] Chalmers, T.C. 'Problems induced by meta-analysis'. *Stat Med*, **10**:971–980, (1991).

[33] Thompson, S.G. and Pocock, S.J. 'Can meta-analysis be trusted?'. *Lancet*, **338**:1127–1130, (1991).

[34] Kassirer, J.P. 'Clinical trials and meta-analysis: What do they do for us'. *New Engl J Med*, **327**:273–274, (1992).

[35] Berlin, J.A., Laird, N.M., Sacks, H.S., and Chalmers, T.C. 'A comparison of statistical methods for combining event rates from clinical trials'. *Stat Med*, **8**:141–151, (1989).

[36] Dickerson, K. and Berlin, J.A. 'Meta-analysis: State-of-the-science'. *Epidemiol Rev*, **14**:154–176, (1992).

[37] Yusuf, S., Wittes, J., Probstfield, J., and Tyroler, H.A. 'Analysis and interpretation of treatment effects in subgroups of patients in randomised clinical trials'. *JAMA*, **266**:81–106, (1991).

[38] DerSimonian, R. and Laird, N.M. 'Meta-analysis in clinical trials'. *Contr Clin Trials*, **7**:177–188, (1986).

[39] Hedges, L.V. and Olkin, I. *Statistical Methods for Meta-Analysis.* Academic Press, San Diago, 1985.

[40] Carlin, J.B. 'Meta-analysis for 2x2 tables: A Bayesian approach'. *Stat Med*, **11**:141–158, (1992).

[41] Skene, A.M. and Wakefield, J.C. 'Hierarchical models for multicentre binary response studies'. *Stat Med*, **9**:919–929, (1990).

[42] Malec, D. and Sedransk, J. 'Bayesian methodology for combining results from different experiments when the specification for pooling are uncertain'. *Biometrika*, **79**:593–601, (1992).

[43] Eddy, D.M., Hasselblad, V., and Shachter, R. 'An introduction to a Bayesian method for meta-analysis: The confidence profile method'. *Medical Decision Making*, **10**:15–23, (1990).

[44] Goodman, S.N. 'Meta-analysis and evidence'. *Contr Clin Trials*, **10**:188–204, (1989).

[45] van Houwelingen, H.C., Zwinderman, K.H., and Stijnen, T. 'A bivariate approach to meta-analysis'. *Stat Med*, **12**:2273–2284, (1993).

[46] Collins, R., Yusuf, S., and Peto, R. 'Overview of randomised trials of diuretics in pregnancy'. *Brit Med J*, **290**:17–23, (1985).

[47] Medical Research Council Working Party. 'MRC trial of treatment of mild hypertension: Principal results'. *Brit Med J*, **291**:97–104, (1985).

[48] Fleiss, J.L. 'The statistical basis of meta-analysis'. *Stat Meth Med Res*, **2**:121–145, (1993).

[49] Mantel, N. and Haenszel, W. 'Statistical aspects of the analysis of data from retrospective studies of disease'. *J Natl Cancer Inst*, **22**:719–748, (1959).

[50] Armitage, P. and Berry, G. *Statistical Methods in Medical Research*. Blackwell Scientific Publications, Oxford, London, Edinburgh, second edition, 1987.

[51] Yusuf, S., Peto, R., Collins, R., and Sleight, P. 'Beta blockade during and after myocardial infarction: An overview of the randomized trials'. *Progress in Cardiovascular Diseases*, **27**:335–371, (1985).

[52] Woolf, B. 'On estimating the relation between blood group and disease'. *Ann Hum Genet*, 19:251–253, (1955).

[53] Rothman, K.J. *Modern Epidemiology.* Little, Brown and Company, Boston, 1986.

[54] Robins, J.M., Breslow, N., and Greenland, S. 'Estimators of the Mantel-Haenszel variance consistent in both sparse data and large strata limiting models'. *Biometrics*, 42:311–323, (1986).

[55] Greenland, S. and Salvan, A. 'Bias in the one-step method for pooling study results'. *Stat Med*, 9:247–252, (1990).

[56] McCullagh, P. and Nelder, J.A. *Generalized Linear Models.* Chapman and Hall, London, 1983.

[57] Longnecker, M.P., Berlin, J.A., Orza, M.J., and Chalmers, T.C. 'A meta-analysis of alcohol consumption in relation to risk of breast cancer'. *JAMA*, 260:652–656, (1988).

[58] Aitkin, M., Anderson, D., Francis, B., and Hinde, J. *Statistical Modelling in GLIM.* Oxford University Press, Oxford, 1989.

[59] Cox, D.R. *The Analysis of Binary Data.* Chapman and Hall, London, 1977.

[60] Gart, J.J. 'Point and interval estimation of the common odds ratio in the combination of 2x2 tables with fixed marginals'. *Biometrika*, 57:471–475, (1970).

[61] Emerson, J.D. 'Combining estimates of the odds ratio: The state of the art'. *Stat Meth Med Res*, 3:157–178, (1994).

[62] Donald, A. and Donner, A. 'A simulation study of the analysis of sets of 2x2 contingency tables under cluster sampling: Estimation of a common odds ratio'. *JASA*, 85:537–543, (1990).

[63] Mehta, C.R. and Walsh, S.J. 'Comparison of exact, mid-P, and Mantel-Haenszel confidence intervals for the common odds ratio across several 2x2 contingency tables'. *American Statistician*, **46**:146–150, (1992).

[64] Antiplatelet Trialists Collaboration. 'Collaborative overview of randomised trials of antiplatelet therapy - I: Prevention of death, myocardial infarction, and stroke by prolonged antiplatelet therapy in various categories of patients'. *Brit Med J*, **308**:81–106, (1994).

[65] Jenicek, M. 'Meta-analysis in medicine: Where we are and where we want to go'. *J Clin Epidemiol*, **42**:35–44, (1989).

[66] Galbraith, R.F. 'A note on graphical presentation of estimated odds ratios from several clinical trials'. *Stat Med*, **7**:889–894, (1988).

[67] Law, M.R. and Thompson, S.G. 'Low serum cholesterol and the risk of cancer: and analysis of the published prospective studies'. *Cancer Causes and Control*, **2**:253–261, (1991).

[68] Pocock, S.J., Smith, M., and Baghurst, P. 'Environmental lead and children's intelligence: a systematic review of epidemiological evidence'. *Brit Med J*, **309**:1189–1197, (1994).

[69] Early Breast Cancer Trialists' Collaborative Group. 'Systematic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy (part 1)'. *Lancet*, **339**:1–15, (1992).

[70] Early Breast Cancer Trialists' Collaborative Group. 'Systematic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy (part 2)'. *Lancet*, **339**:71–85, (1992).

[71] Antiplatelets Trialists' Collaboration. 'Secondary prevention of vascular disease by prolonged antiplatelet treatment'. *Brit Med J*, **296**:320–331, 1988).

[72] Bailey, K.R. 'Inter-study differences: How should they influence the interpretation and analysis of results'. *Stat Med*, 6:351–358, (1987).

[73] Peto, R. Discussion. *Stat Med*, 6:349–350, (1987).

[74] Meier, P. 'Meta-analysis of clinical trials as a scientific discipline (commentary)'. *Stat Med*, 6:329–331, (1987).

[75] Raghunathan, T.E. and Yoichi II. 'Analysis of binary data from a multicentre clinical trial'. *Biometrika*, 80:127–139, (1993).

[76] Spector, T.D. and Thompson, S.G. 'The potential and limitations of meta-analysis'. *J Epidemiol and Community Health*, 45:89–92, (1991).

[77] Maritz, J.S. and Lwin, T. *Empirical Bayes Methods.* Chapman and Hall, London, New York, 1989.

[78] Becker, R.A., Chambers, J.M., and Wilks, A.R. *The New S Language. A Programming Environment for Data Analysis and Graphics.* Wadsworth & Brooks/Cole, Pacific Grove, California, 1988.

[79] Cox, D.R. and Hinkley, D.V. *Theoretical Statistics.* Chapman and Hall, London, 1990.

[80] Steering Committee of the Physicians' Health Study Research Group. 'Final report of the aspirin component of the ongoing Physicians' Health Study'. *New Engl J Med*, 321:129–135, (1989).

[81] Rosner, B. 'A generalization of the paired t-test'. *Appl Statist*, 31:9–13, (1982).

[82] Edwards, A.W.F. *Likelihood.* Cambridge University Press, Cambridge, 1972.

[83] Clayton, D. and Hills, M. *Statistical Models in Epidemiology.* Oxford University Press, Oxford, 1994.

[84] Laird, N.M. 'Nonparametric maximum likelihood estimation of a mixing distribution'. *JASA*, **73**:805–811, (1978).

[85] Dempster, A.P., Laird, N.M., and Rubin, D.B. 'Maximum likelihood from incomplete data via the EM algorithm'. *J Roy Statist Soc, Series B*, **39**:1–38, (1977).

[86] Agresti, A. *Categorical Data Analysis.* John Wiley and Sons, New York, 1990.

[87] Cytel Software Corporation, Cambridge, USA. *StatXact: User Manual (Version 2)*, 1991.

[88] Zelen, M. 'The analysis of several 2x2 contingency tables'. *Biometrika*, **58**:129–137, (1971).

[89] Mehta, C.R., Patel, N.R., and Gray, R. 'Computing an exact confidence interval for the common odds ratio in several 2x2 contingency tables'. *JASA*, **80**:969–973, (1985).

[90] Colditz, G.A., Brewer, T.F., Berkey, C.S., Burdick, E., Fineberg, H.V., and Mosteller, F. 'Efficacy of BCG vaccine in the prevention of tuberculosis: Meta-analysis of the published literature'. *JAMA*, **271**:698–702, (1994).

[91] Morris, C.N. 'Parametric empirical Bayes inference: Theory and applications'. *JASA*, **78**:47–65, (1983).

[92] DuMouchel, W.H. and Harris, J.E. 'Bayes methods for combining the results of cancer studies for humans and other species'. *JASA*, **78**:293–307, (1983).

[93] Rubin, D.B. 'Estimation in parallel randomized experiments'. *Journal of Educational Statistics*, **6**:377–401, (1981).

[94] Geman, S. and Geman, D. 'Stochastic relaxation, Gibbs distributions and the bayesian restoration of images'. *IEEE Transactions on Patterns Analysis and Machine Intelligence*, **6**:721–741, (1984).

[95] Naylor, J.C. and Smith, A.F.M. 'Applications of a method for the efficient computation of posterior distributions'. *Appl Statist*, **31**:214–225, (1982).

[96] Smith, A.F.M., Skene, A.M., Shaw, J.E.H., Naylor, J.C., and Dransfield, M. 'The implementation of the Bayesian paradigm'. *Communications in Statistics-Theory and Methods*, **14**:1079–1102, (1985).

[97] Eddy, D.M., Hasselblad, V., and Shachter, R. *Meta-Analysis by the Confidence Profile Method: The Statistical Synthesis of Evidence.* Academic Press Inc., USA, 1992.

[98] Gelfand, A.E. and Smith, A.F.M. 'Sampling-based approaches to calculating marginal densities'. *JASA*, **85**:398–409, (1990).

[99] Gelfand, A.E., Hills, S.E., Racine-Poon, A., and Smith, A.F.M. 'Illustration of Bayesian inference in normal data models using Gibbs sampling'. *JASA*, **85**:972–985, (1990).

[100] Matthews, J.N.S. 'A refinement to the analysis of serial data using summary measures'. *Stat Med*, **12**:27–37, (1993).

[101] SAS Institute Inc., Cary, NC. *SAS technical report P-229, release 6.07, The MIXED procedure (pp 287-366)*, 1992.

[102] Blom, G. *Statistical Estimates and transformed Beta variables.* John Wiley, New York, 1958.

[103] Shapiro, S.S. and Francia, R.S. 'An approximate analysis of variance test for normality'. *JASA*, **67**:215–216, (1972).

[104] Royston, P. 'A toolkit for testing for non-normality in complete and censored samples'. *The Statistician*, **42**:37–43, (1993).

[105] Shapiro, S.S. and Wilk, M.B. 'An analysis of variance test for normality (complete samples)'. *Biometrika*, **52**:591–611, (1965).

[106] Stephens, M.A. 'EDF statistics for goodness of fit and some comparisons'. *JASA*, **69**:730–737, (1974).

[107] Sinclair, C.D. and Spurr, B.D. 'Approximation to the distribution function of the Anderson-Darling statistic'. *JASA*, **83**:1190–1191, (1988).

[108] Dempster, A.P. and Ryan, L.M. 'Weighted normal plots'. *JASA*, **80**:845–850, (1985).

[109] Jones, M.P., O'Gorman, T.W., Lemke, J.H., and Woolson, R.F. 'A Monte Carlo investigation of homogeneity tests of the odds ratio under various sample size configurations'. *Biometrics*, **45**:171–181, (1989).

[110] Ewertz, M., Duffy, S.W., Adami, H-O., Kvåle, G., Lund, E., Meirik, O., Mellemgaard, A., Soiri, I., and Tulinius, H. 'Age at first birth, parity and risk of breast cancer: A meta-analysis of 8 studies from Nordic countries'. *Int J Cancer*, **46**:597–603, (1990).

[111] Breslow, N.E. and Day, N.E. *Statistical Methods in Medical Research, I. The Analysis of Case-control Studies.* International Agency for research on Cancer, Lyon, 1980.

[112] Family Heart Study Group. 'Randomised controlled trial evaluating cardiovascular screening and intervention in general practice: Principal results of British family heart study'. *Brit Med J*, **308**:313–320, (1994).

[113] Raudenbush, S.J., Becker, B.J., and Kalaian, H. 'Modelling multivariate effect sizes'. *Psychol Bull*, **103**:111–120, (1988 .

[114] Pocock, S.J., Geller, N.L., and Tsiatis, A.A. 'The analysis of multiple endpoints in clinical trials'. *Biometrics*, **43**:487–498, (1987).

[115] Miller, R.G. *Simultaneous Statistical Inferences.* Springer-Verlag, New York, 1981.

[116] Armitage, P. and Parmer, M. Some approaches to the problem of multiplicity in clinical trials. In *Proceedings of the XIIIth International Biometric Conference*, 1986.

[117] Iaffaldano, M.T. and Muchinsky, P.M. 'Job satisfaction and job performance: A meta-analysis'. *Psychol Bull*, **97**:251–273, (1985).

[118] Miller, R.C. and Berman, J.S. 'The efficacy of cognitive behaviour therapies: A quantitative review of the research evidence'. *Psychol Bull*, **94**:39–53, (1983).

[119] Rosenthal, R. and Rubin, D.B. 'Meta-analytic procedures for combining studies with multiple effect sizes'. *Psychol Bull*, **99**:400–406, (1983).

[120] O'Brien, P.C. 'Procedures for comparing samples with multiple endpoints'. *Biometrics*, **40**:1079–1087, (1984).

[121] H. Goldstein. *Multilevel Models in Educational and Social Research.* Griffin, London, 1987.

[122] Donald, A. and Donner, A. 'Adjustments to the Mantel-Haenszel chi-square statistic and odds ratio variance estimator when the data are clustered'. *Stat Med*, **6**:491–499, (1987).

[123] Donner, A. and Donald, A. 'Analysis of data arising from a stratified design with the cluster as unit of randomization'. *Stat Med*, **6**:43–52, (1987).

[124] Donner, A. 'The analysis of intraclass correlation in multiple samples'. *Ann Hum Genet*, **49**:75–82, (1985).

[125] Donner, A and Hauck, W. 'Estimation of a common odds ratio in paired-cluster randomization designs'. *Stat Med*, **31**:599–607, (1989).

[126] Donner, A., April 18, 1994. Personal communication.

[127] Freund, J.E. and Walpole R.E. *Mathematical Statistics.* Prentice-Hall International, New Jersey, fourth edition, 1987.

[128] Gail, M.H., Byar, D.P., Pechacek, T.F, and Corle, D.K. 'Aspects of statistical design for community intervention trial for smoking cessation (COMMIT)'. *Contr Clin Trials*, **13**:6–21, (1992).

[129] Fisher, R.A. *The Design of Experiments.* Hafner, New York, eighth edition, 1966.

[130] Donner, A. and Klar, N. 'Confidence interval construction for effect measures arising from cluster randomization trials'. *J Clin Epidemiol*, 46:123–131, (1993).

[131] Korn, E.L. 'The paired t-test'. *Appl Statist*, **33**:230–231, (1984). Letter to editor.

[132] Robins, J., Greenland, S., and Breslow, N.E. 'A general estimator for the variance of the Mantel-Haenszel odds ratio'. *Am J Epidemiol*, **124**:719–723, (1986).